

**ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA INFORMÁTICA**

**GRADO EN INGENIERÍA INFORMÁTICA**

**Curso Académico 2022/2023**

**Trabajo Fin de Grado**

**INGENIERÍA DE DATOS CON EL FRAMEWORK DE BIG DATA SPARK Y SCALA**

**Autor**: Álvaro Sánchez Pérez

**Directores**: Juan Manuel Serrano Hidalgo

Índice

[1. Introducción 3](#_Toc118635906)

[2. Objetivos 3](#_Toc118635907)

[3. Descripción informática 4](#_Toc118635908)

[1. Fuentes de datos 4](#_Toc118635909)

[1.1. Obtención de los datos 8](#_Toc118635910)

[2. Programación de queries 10](#_Toc118635911)

[3. Visualización de queries 17](#_Toc118635912)

[4. Despliegue en AWS EMR 22](#_Toc118635913)

[2. Experimentos / validación 22](#_Toc118635914)

[2.1. Consultas realizadas 22](#_Toc118635915)

[2.2. Análisis de requisitos no funcionales 22](#_Toc118635916)

[3. Conclusiones 22](#_Toc118635917)

[4. Bibliografía 22](#_Toc118635918)

[5. Apéndices 22](#_Toc118635919)

# Introducción

Prueba

# Objetivos

# Descripción informática

## Fuentes de datos

Los datos meteorológicos, han sido obtenidos a través de la página de *AEMET OpenData*, se tratan tanto de datos que muestran la información de forma mensual como de forma diaria desde el año 2007 hasta el año actual 2022.

Imagen que contiene Interfaz de usuario gráfica

Descripción generada automáticamente

Resaltar que el diagrama mostrado, aparecen los id de las columnas y valores que deberían de encontrase en estas mediante los datos obtenidos, ya que así nos lo indican los metadatos que nos proporcionan. Pero en cambio tanto el nombre, como la provincia no aparecen en la mayoría de los archivos de datos que nos proporcionan, y por ello debemos de hacer uso de la tabla *AEMET ID*,que gracias a esta podremos relacionar mediante el indicativo de que provincia y ubicación se trata.

La estructura en la que vienen estos datos se puede observar en el anterior diagrama, todos los valores vienen en formato String, los cuales transformaremos mediante el casteo correspondiente gracias a Spark, para así poder trabajar posteriormente con ellos de una manera más cómoda. Deberemos de realizar lo siguiente para la transformación.

1. Primero creamos el esquema correspondiente a los datos, en este ejemplo lo realizaremos con los datos por día. En este esquema, le indicaremos tanto en nombre de la columna, como el tipo de dato que se trata, en nuestro caso todos String, y por último si la columna puede contener valores nulos.

Texto

Descripción generada automáticamente

1. A continuación, realizaremos la lectura de los datos y casteo de los datos.

Texto, Carta

Descripción generada automáticamente

* 1. Para la lectura le proporcionaremos el esquema creado anteriormente en la variable *schema*, y a su vez le indicamos en la ubicación donde se encuentran los datos, además del formato de fichero.



* 1. Finalmente, modificaremos el tipo de dato de las columnas que lo requieran. Usaremos *withColumn* donde primero le indicaremos el nuevo nombre de la columna que nos generará, en este caso usaremos el mismo nombre de la comuna que vamos a modificar para que así ocurra una sustitución de esta, y en la segunda parte será donde le indiquemos la columna con la que queremos trabajar y el cambio que deseamos.



He de mencionar que, los datos que se tuvieron que transformar a tipo Double se debió realizar una modificación previa, sustituyendo las comas por puntos, ya que de otra manera no resultaba posible realizar el casteo debido a que para que se considere un numero decimal debe de venir separada la parte entera de la decimal mediante un punto.



1. Por último, guardaremos los datos en formato Parquet, particionados mediante el campo *indicativo.*



Quedando particionado en memoria de la siguiente manera:

Texto

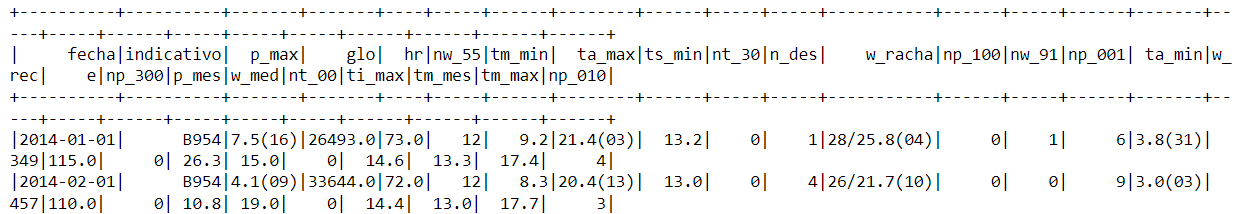
Descripción generada automáticamente con confianza media

Y dentro de cada carpeta algo similar a lo siguiente:

Imagen que contiene Texto

Descripción generada automáticamente

Respecto a los datos mensuales nos quedaríamos únicamente con las siguientes columnas.



De las que podríamos destacar las siguientes: *tm\_mes* la cual nos muestra una temperatura media mensual de la ubicación, *tm\_max* y *tm\_min* siendo las temperaturas medias máximas y mínimas respectivamente, *ta\_max* y *ta\_min* como las temperaturas máximas y mínimas absolutas del mes y *p\_mes* la cual nos indica precipitación total en ese mes. En caso de querer obtener información acerca de alguna otra variable, se puede acceder a esta de las siguientes maneras: a través del fichero *metadataMonth.json* incluido en la carpeta *data* del proyecto o realizando cualquier consulta sobre climatologías mensuales/anuales en la página de *AEMET OpenData* donde obtendrá un enlace acerca de los metadatos donde encontrar este tipo de información.

En relación con los datos diarios, tendríamos lo siguiente:

Calendario

Descripción generada automáticamente con confianza baja

En este caso nos hemos quedado con todas las columnas, de las que resaltaríamos las siguientes: *tmed* la cual muestra la temperatura media diaria, *prec* que nos ofrece información acerca de la precipitación diaria, tmax *y* *tmin* las cuales nos muestran la temperatura máxima y mínima diaria respectivamente, y *horatmax* y *horatmin* que, de forma correspondiente, presentan la hora y minuto de latemperatura máxima y mínima. En caso de querer obtener información acerca de alguna otra variable, se puede acceder a esta de las siguientes maneras: a través del fichero *metadataDay.json* incluido en la carpeta *data* del proyecto o realizando cualquier consulta sobre climatologías diarias en la página de *AEMET OpenData* donde obtendrá un enlace acerca de los metadatos donde encontrar este tipo de información

He de mencionar que, a la hora de descargar los datos se encuentran en un formato JSON multilínea, el cual no llega a ser lo suficientemente óptimo para la lectura con Apache Spark. Por lo tanto, se realizaron cambios en todos los ficheros transformándolos a ficheros JSON que contuvieran toda la información en una única línea, esto se llevó a cabo mediante el siguiente comando utilizando la consola de Windows: **FOR %a IN (../data/\*.json) DO jq . -c "%a" > "../JSONLine/%a”**. También cabe destacar que los datos una vez leídos en este formato JSON, la mayoría fueron transformados al formato Parquet, particionados por el indicativo de cada estación, ya que resulta más óptimo para la realización de consultas con Spark.

## Obtención de los datos

Para la obtención de los datos, se crearon unos pequeños programas en el lenguaje Java. Estos programas, se basan en la simulación de los pasos que deberíamos de seguir para la descarga y los podemos encontrar en las carpetas *Desca*rgaDatosPorDias y *DescargaDatosPorMeses*. Para la simulación de estos pasos, se hizo uso de la librería *Selenium*, a través de la cual se puede manejar un navegador web pudiendo realizar diversas acciones sobre la página mostrada. Por lo tanto, de manera resumida, los pasos que realizarían ambos programas para la descarga serían los siguientes.

Texto

Descripción generada automáticamente

Abriríamos un nuevo navegador donde accederemos al sitio web de *AEMET OpenData*. En nuestro caso le pasamos la URL del sitio web a través de la variable *baseURL*.

Texto

Descripción generada automáticamente

Una vez dentro de esta página, el primer paso sería introducir la clave API correspondiente a nuestro usuario, está la podremos solicitar también a través de la misma página. Posteriormente, dependiendo del tipo de información que quisiéramos obtener, ya fuera información por días o por meses, buscaríamos los desplegables y elementos necesarios mediante su *id* o *xpath*.

Después de seleccionar la información deseada, se nos abriría una nueva página donde nos proporciona diferentes tipos de información, además del enlace de la página donde se encontrarán los datos deseados.

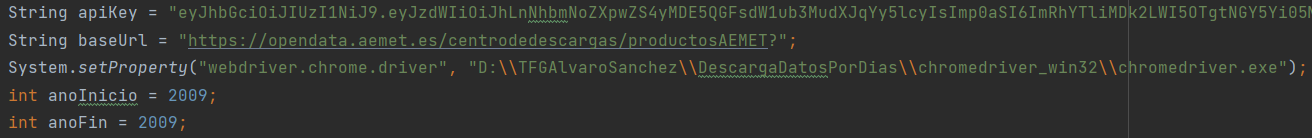


En caso de que la consulta a los datos correspondientes se haya realizado con éxito, accederemos al nuevo enlace que nos muestra, donde obtendremos la información y la guardaremos en un nuevo fichero con el nombre correspondiente a la consulta.



Por último, cerramos todas las pestañas y navegadores que se hubieran abierto, y procederíamos a realizar los mismos pasos con otro rango de fechas o en otra estación meteorológica, el objetivo es obtener la información de todas las estaciones existentes entre el rango de fechas establecido.

He de destacar también, que al comienzo del código existen una serie de variables que pueden ser cambiadas por los usuarios. Desde introducir su correspondiente clave API, cambiar el controlador de Chrome en caso de que se esté usando una versión diferente del navegador, o cambiar las fechas en caso de que se quisieran obtener datos en un rango diferente.



## Programación de queries

**Año de la temperatura máxima promedio en cada mes**

En esta consulta, realizaremos una comparación del mismo mes en diferentes años, para saber en qué año se registró la temperatura máxima promedia entre todas las estaciones en este mes. Por último, lo representaremos tanto mediante una tabla, como de manera gráfica para poder observar las diferencias de manera visual entre los distintos años.

Para realizar esta consulta utilizaremos los datos meteorológicos mensuales, ya que nos ofrecen la temperatura máxima registrada en cada mes de una manera bastante sencilla. Por lo tanto, el primer paso que deberíamos de realizar sería la lectura de estos datos.

A continuación, deberíamos de agrupar estos datos por fecha, ya que, al realizar la lectura, tenemos la información de muchas estaciones meteorológicas diferentes en el mismo mes y año, por lo tanto, los agruparemos haciendo que coincida la fecha, y realizando una media de la temperatura máxima en todas las estaciones. Por último, dividiremos los datos por meses y ordenándolos de manera descendente a través de nuestro objetivo, la temperatura máxima, ya que así estaríamos consiguiendo obtener el año, en el cual este mes fue el más caluroso.

**Olas de calor**

En esta consulta, desearemos obtener las diferentes olas de calor que han producido cada año en España, con el objetivo de observar si es verdad que se está produciendo un aumento de las temperaturas durante los últimos años. Para ello, calcularemos las olas de calor teniendo en cuenta la temperatura máxima diaria en cada estación meteorológica, y realizaremos los cálculos necesarios para finalmente mostrar los resultados sobre un mapa donde el usuario podrá elegir el año del cual desea obtener la información. Resaltar que, se considerará ola de calor cuando la temperatura máxima supere o iguale los 40 grados durante más de tres días consecutivos.

Debido a la exactitud que necesitamos para esta consulta se utilizarán los datos meteorológicos diarios, ya que requeriremos los datos de las temperaturas máximas con una gran exactitud en el marco temporal. Por lo tanto, el primer paso para comenzar con esta consulta sería realizar la lectura de estos datos.

Una vez leídos los datos, deberemos de eliminar aquellos donde no se llegue a alcanzar la temperatura establecida, en nuestro caso eliminaremos aquellos que tengan una temperatura menor a 40 grados. Posteriormente, deberemos de identificar posibles olas de calor, aun sin tener en cuenta la duración de estas, para ello dividiremos los datos por estación meteorológica y año, y a cada fecha se le asignara un identificador, en caso de que las fechas que sean consecutivas se les asignara el mismo identificador. Por ir resumiendo, ahora mismo lo que tendríamos serían los días que presentan temperaturas mayores o iguales a 40 grados, donde además se les habrá asignado un identificador, con el que posteriormente podremos determinar la duración de la ola de calor.

A continuación, agruparemos los datos por el identificador asignado a cada fecha, donde contaremos el número de apariciones de cada identificador, y por último eliminaremos aquellos no válidos, quedándonos únicamente con los que se encuentran más de 3 veces. Ya que como se comentó anteriormente, gracias al número de veces que se encuentre un identificador podremos determinar la duración de la ola de calor.

Texto

Descripción generada automáticamente

Texto

Descripción generada automáticamente

La consulta se podría dividir en dos partes, una primera donde calculamos las diferentes olas de calor, y una segunda donde se agruparían estas olas de calor por provincias y obtenemos la información necesaria para representarla posteriormente en forma de mapa.

El primer paso como se indicó anteriormente sería la lectura de los diferentes datos. Primero de ello, leeremos los datos meteorológicos, los cuales guardaremos en la variable *data*. Podemos observar únicamente 3 columnas ya que, al tratarse de una lectura de tipo Parquet, Spark únicamente lee las columnas que va a utilizar.

Texto, Carta

Descripción generada automáticamente

También leeremos el fichero a través del cual relacionaremos el indicativo de las estaciones con su provincia y ubicación. Estos datos los guardaremos en la variable *stations* y contendrá una información similar a la siguiente.

Tabla

Descripción generada automáticamente con confianza baja



Además, al comienzo de la consulta nos crearemos nuestra función ventana en la variable *window*. La utilizaremos posteriormente puesto que, a través de esta dividiremos los datos mediante el año y su estación meteorológica, y a su vez ordenaremos estas particiones mediante la fecha de manera ascendente.



Seguidamente comenzamos con la primera parte de la consulta donde obtendremos las distintas olas de calor en cada estación, guardaremos esta información en la variable *results*. El primero de los pasos será eliminar las filas que no utilizaremos, para ello realizamos un filtro en la columna *tmax*, la cual nos proporciona la información de la temperatura máxima, por ello eliminaremos aquellas filas que no contengan ningún tipo de información acerca de esta, y aquellas que no superen la temperatura mínima establecida.



A continuación, haremos uso de la función de Spark *year* con la cual obtendremos el año a través de la fecha. Guardaremos esta nueva información en una nueva columna a la que llamaremos *año*.



A su vez, también crearemos una nueva columna a la cual llamaremos *n\_fila*, en esta le asignaremos un número a cada fila utilizando la función *row\_number*, la cual a su vez hará uso de la función ventana que hemos creado anteriormente. Mediante la función *row\_number*, asignaremos valores de manera secuencial comenzando desde el 1. Y gracias a la función ventana, estamos consiguiendo asignar un numero de fila que va irá en aumento dependiendo de la fecha ya que era el orden que le habíamos establecido. A su vez cada vez que se trate de un nuevo año o de una nueva estación comenzará a contar de nuevo desde el 1, ya que eran las divisiones que le habíamos indicado cuando nos creamos la función. Quedándonos la consulta por el momento de la siguiente manera.

Tabla

Descripción generada automáticamente



Continuaremos creándonos una nueva columna llamada id, en esta realizaremos una resta de la columna *fecha* con el *n\_fila*. De esta manera estaríamos consiguiendo identificar las distintas olas de calor, ya que en caso de tratarse de días consecutivos en la nueva columna *id* contendrán el mismo valor. Por ejemplo, en la anterior imagen tenemos estas filas:

Texto

Descripción generada automáticamente

Como podemos observar se tratan de fechas consecutivas, por lo tanto, si realizamos la resta de la columna *fecha* – *n\_fila* nos quedaría lo siguiente en ambas:

Texto

Descripción generada automáticamente con confianza media

Como podemos, observar tendrían el mismo valor en la columna id y gracias a este estaríamos identificando días consecutivos con temperaturas lo suficientemente altas. Por lo tanto, este proceso lo realizamos con todos nuestros datos, y tendríamos identificadas las distintas posibles olas de calor, ya que aun deberemos tener él cuenta el número de días que duran para que se consideren validas.



Retomando la consulta, ahora agruparemos las distintas posibles olas de calor, para ello realizaremos *groupBy* con las columnas *indicativo*, *año* e *id*. Posteriormente, mediante la función *agg*, llamaremos a diferentes funciones Spark a través de las cuales obtendremos información. Por ejemplo, mediante la función *count* con la columna *id*, obtendremos la duración en días de las altas temperaturas, ya que estaremos contando el número de veces que podemos encontrar ese mismo *id* en la misma estación meteorológica y año, a su vez, haciendo uso de la función *alias* le establecemos el nombre a la nueva columna que se nos generará. Mediante las funciones *avg*, *max* y *min* obtendremos la temperatura media, máxima y mínima respectivamente acerca de esa posible ola de calor.



Para ir finalizando con la primera parte de la consulta, realizamos un nuevo filtrado de los datos, en este caso quedándonos con lo que nosotros hemos considerado olas de calor, aquellas cuya duración es mayor a 3 días.



Realizamos una unión mediante la función *join* con nuestros datos guardados en la variable *stations*, la cual contenía información extra acerca de las estaciones. Estos datos tenían tres columnas, *provincia*, *indicativo* y *ubicación*, por lo tanto, los unimos mediante la columna común a ambas tablas, *indicativo*.



Por último, elegimos las columnas relevantes para nuestro resultado y mediante la función *round* redondeamos el valor perteneciente a la columna *avg(tmax)*, estableciéndole que contenga únicamente dos decimales.

Ahora mismo, tendríamos en la variable *results* las distintas olas de calor en cada estación meteorológica, viéndose de la siguiente manera:

Tabla

Descripción generada automáticamente

Donde la columna *dias* representaría la duración de esa ola de calor, *avg(tmax)* la temperatura media, *max(tmax)* y *min(tmax)* las temperaturas máximas y mínimas respectivamente.

En la segunda parte de la consulta, preparamos los datos para su representación en el mapa provincial de España, además de guardarlos en un formato Parquet para su posterior lectura en Python, ya que actualmente en Scala no es posible realizar una representación sobre un mapa.



Como primer paso, nos creamos una variable que hemos llamado *resultsSave*, en ella se encontrarán los datos agrupados por provincias preparados para su posterior representación. Por lo tanto, el siguiente paso a realizar seria agrupar los datos por provincia y año mediante la función *groupBy*. Mediante la función *agg* ejecutaremos diferentes funciones, por medio de la función *count* obtendremos el número de olas de calor en un mismo año en una provincia, ya que gracias a contar número de veces que aparece una provincia en los anteriores resultados podremos determinar este valor y mediante las funciones *avg*, *max* y *min* obtendremos de manera respectiva la temperatura media, máxima y minina para esa provincia, con respecto a las olas de calor pertenecientes a un año.

Texto

Descripción generada automáticamente con confianza media

Por último, seleccionamos las columnas deseadas y cambiamos el nombre de algunas de ellas mediante la función *alias*.

Imagen que contiene Texto

Descripción generada automáticamente

Para guardar los datos en formato Parquet el nombre de las columnas no puede contener ningún espacio en blanco, por lo tanto, deberemos de darles un nuevo nombre a estas columnas. Lo realizamos a través de la función *withColumnRenamed* donde primero le indicamos el nombre actual de la columna, y a continuación el nuevo nombre que le queremos establecer.

## Visualización de queries

**Año de la temperatura máxima promedio en cada mes**

Texto

Descripción generada automáticamente

Como se puede observar,tal y como se describió anteriormente, el primer paso que realizamos es la lectura de los datos, los cuales los estamos leyendo en un formato Parquet para que resulte más óptimo y además estamos eliminando las filas que contienen valores nulos mediante la función *na.drop()*, la lectura de estos datos la almacenamos en la variable *data*. Al realizar la lectura de estos, se encontrarían de la siguiente manera.

Texto, Carta

Descripción generada automáticamente



Por otro lado, también nos creamos la función ventana en la variable *window* que utilizaremos posteriormente. Le indicaremos la manera de la cual deseamos que se dividan los datos, en nuestro caso será a través de los meses y a su vez le señalamos la manera en la que se deberán de ordenar, teniendo en cuenta la temperatura máxima.



A continuación, procedemos a realizar la consulta que almacenaremos en la variable *dataWindow*. El primer paso, que realizamos con los datos es la modificación de la columna de temperatura máxima, ya que esta nos viene con el siguiente formato temperaturaMáximaAlcanzada(díaDelMesEnElQueSeAlcanzó) por lo que Spark estaría tratando esta columna como un String, cuando nuestro objetivo es manejarlo como un tipo Double. Para esto realizamos lo siguiente, separamos mediante la función *Split* la temperatura máxima del día en que se alcanzó, y nos quedamos únicamente con la temperatura máxima, por último, realizamos un casteo al tipo de datos que deseamos.



El segundo paso, de la consulta seria agrupar los datos de las diferentes estaciones por fecha, para ello utilizamos la función *groupBy* indicándole la columna a través de la cual deseamos realizar la agrupación. Mediante la función *agg* y *avg*, obtenemos y calculamos la media de la temperatura máxima de las filas que hemos agrupado. Actualmente tendríamos el DataFrame que se muestra a continuación, que contendría únicamente las fechas y sus respectivas temperaturas.

Texto, Carta

Descripción generada automáticamente



En el siguiente paso, obtendremos tanto el mes y el año, a partir de la fecha. Para esto utilizamos las funciones *month* y *year* con las cuales obtenemos tanto el mes y el año respectivamente. Mediante la función *select* seleccionaremos únicamente las columnas que vamos a necesitar.



Para ir finalizando, añadimos una columna con el valor que nos devuelve la función *dense\_rank*,que hace uso de la ventana *window* anteriormente definida. Mediante esta función, obtenemos la posición en la que se encuentra un valor teniendo en cuenta la expresión *orderBy* establecida en la función ventana (*window*). Por lo tanto, en nuestro nos devolverá una clasificación de los años más calurosos dividido por meses, ya que era la partición que se estableció en la función *window*. Quedando la consulta de la siguiente manera después de aplicar la función ventana.

Tabla

Descripción generada automáticamente



Por último, realizamos un filtrado quedándonos únicamente con las filas que poseen en la columna *dense\_rank* un valor igual a 1, obteniendo así el año en el que se obtuvo la mayor temperatura máxima en un determinado mes.

Texto

Descripción generada automáticamente

Se implementa también la función *countByYear*, la cual nos facilitará los datos que utilizaremos para la representación gráfica, ya que nos devuelve la cantidad de meses por cada año que se encuentran en la consulta realizada anteriormente.

Para finalizar ofrecemos los resultados tanto en forma de tabla como de una forma gráfica que sería para la cual habríamos hecho uso de la función de la *countByYear*.

Tabla

Descripción generada automáticamente

Gráfico, Gráfico de barras

Descripción generada automáticamente

**Olas de calor**

Texto

Descripción generada automáticamente

Podremos observar los resultados de diferentes maneras, una de ellas sería mostrar las diferentes olas de calor a través de una tabla, donde se nos muestre toda la información de manera detalla, ordenada por ubicación y año:

Tabla

Descripción generada automáticamente con confianza media

Texto

Descripción generada automáticamente con confianza baja

Otra manera sería la siguiente, con la que podremos observar el numero de olas de calor en los diferentes años. Como se puede ver en los ultimos años estas han aumentado de una manera considerable.

Tabla

Descripción generada automáticamente

Y por ultimo, una representación mediante un mapa, para la cual fue necesario guardar los datos de los resultados, ya que en Scala aun no existen librerias que nos permitan este tipo de representaciones, por lo tanto haremos uso de Python.

Para esta representacion, haremos uso de las siguientes librerias en Python:

Una captura de pantalla de un celular con texto e imagen

Descripción generada automáticamente con confianza media

Utilizaremos la librería *json* para leer el fichero geoJSON, el cual es un fichero que nos ofrece datos geofraficos con el cual representaremos las diferentes pronvincias españolas. La librería *pandas* la usaremos para la lectura y manipulacion de los datos, mediante *ploty* podremos representar nuestro gráfico y por ultimo utilizaremos *ipywidgets* para poder utilizar elementos interactivos de HTML en Jupyter notebook.

## Despliegue en AWS EMR

# Experimentos / validación

## Consultas realizadas

## Análisis de requisitos no funcionales

# Conclusiones

# Bibliografía

# Apéndices