MÓDULO

EJERCICIO FINAL PROGRAMACIÓN BIG DATA

EJERCICIO FINAL

EJERCICIO FINAL

Introducción.

En este ejercicio final vas a tener que aplicar los conocimientos que has adquirido a lo largo de las unidades que componen el segundo módulo de Tecnologías Big Data.

Además, también vas a tener que hacer algo de investigación y aprender a utilizar una plataforma real en la nube para resolver el ejercicio.

Se trata de que demuestres no solo los conocimientos, si no también tu capacidad para trabajar con estas tecnologías.

¡No te preocupes, es más sencillo de lo que puedas pensar!

Objetivos y entregables

Para completar este ejercicio tendrás que utilizar varios ficheros anexos al ejercicio. Por un lado, encontrarás un archivo .zip que contiene los ficheros de datos a utilizar. Descomprime el contenido en una carpeta de tu ordenador. El archivo consta de tres ficheros:

- precio_md.csv : datos de precios horarios de electricidad
- preveol.csv : datos de generación eólica prevista para cada hora
- prvdem.csv : datos de demanda prevista para cada hora

Si quieres, puedes ver el contenido usando Excel o un editor de texto. O mejor, puedes usar R o Python para familiarizarte con los datos antes de empezar. ¡Solamente recuerda no sobrescribir el contenido de los ficheros!

Por otro lado, para la segunda parte del ejercicio tendrás que utilizar un fichero con extensión .dbc que también encontrarás anexo al ejercicio. Se trata de un cuaderno de trabajo de Databricks que deberás cargar siguiendo las instrucciones que te indicamos más adelante. Contiene una serie de tareas de programación con Spark y Pandas que deberás completar en el propio cuaderno. Cuando las hayas completado, tendrás que exportarlo y enviarlo como resolución del ejercicio.

A lo largo del enunciado encontrarás las indicaciones adecuadas. Si necesitas alguna aclaración, puedes consultar a los profesores del curso.

1. DESARROLLO DEL EJERCICIO – USANDO DATABRICKS

Para realizar el ejercicio final vas a tener que usar una plataforma en la nube: Databricks.

Databricks es una plataforma de análisis de datos en la nube basada en Apache Spark. De hecho, Databricks fue fundada por los creadores de Spark.

Databricks permite a los usuarios cargar ficheros de datos en la nube y crear cuadernos de trabajo interactivos (notebooks, como en Jupyter) para analizar los datos usando Spark y compartir fácilmente nuestro trabajo y resultados con otras personas.

1.1. CREAR UNA CUENTA DE USUARIO

Databricks nos permite crear una cuenta gratuita para acceder a su versión "Community". No tiene todas las funcionalidades de las cuentas de pago, pero incluye lo fundamental para trabajar con Spark en la nube. Y como he dicho, es gratis.

Lo primero que tienes que hacer es crearte una cuenta. Puedes hacerlo a través de este enlace:

https://databricks.com/try-databricks

Elige la opción Community Edition y pulsa en Get Started. Rellena los campos del formulario y sigue las instrucciones.

Una vez que tengas activada tu cuenta, podrás acceder a Databricks a través de este enlace:

https://community.cloud.databricks.com

Al entrar, verás la pantalla principal desde donde se accede a las distintas funcionalidades:

Desde la pantalla principal puedes crear nuevos cuadernos de trabajo, importar datos a tu cuenta o crear un nuevo cluster de Spark con el que trabajar.

También dispones de enlace a tutoriales básicos (Explore the Quickstart Tutorial) o a la documentación de Databricks y Spark.

1.2. CREAR UN CLUSTER EN DATABRICKS.

Para empezar a trabajar, lo primero que necesitarás es un cluster de Spark.

En la barra de navegación de la izquierda, pulsa en el botón “Clusters”

Accederás a la ventana de creación de clusters. En la edición “Community” gratuita no podemos seleccionar muchas opciones. Simplemente dale un nombre al cluster, selecciona la última versión de Databricks disponible (normalmente ya vendrá elegida por defecto) y usar Python versión 3.

Cuando ya lo tengas, pulsa el botón “Create Cluster”

A continuación, podrás ver en la ventana de Clusters una nueva entrada en tabla de clusters interactivos correspondiente al que acabas de crear. Verás que durante unos segundos (hasta unos pocos minutos) aparece en estado “Pending”. Eso significa que Databricks está configurando y arrancando el cluster.

Una vez que haya completado este proceso, el cluster aparecerá en estado “Running”. Eso significa que ya puedes utilizarlo para ejecutar tus notebooks.

No es necesario que crees un cluster cada vez que entres en tu cuenta. Databricks guarda la configuración del cluster, de manera que la siguiente vez solamente tengas que seleccionarlo para que arranque y usarlo en tus cuadernos de trabajo.

1.3. CARGAR DATOS.

El siguiente paso consiste en cargar nuestros datos para poder analizarlos.

Vuelve a la ventana principal de Databricks pulsando en el botón “databricks” de la barra de la izquierda

En el panel central, elige la opción “Import & Explore Data”.

Aparecerá una nueva ventana desde donde podrás subir nuevos ficheros de datos:

Mantén seleccionada como fuente de datos (“Data source”) la opción “Upload File”.

En el cuadro inferior podrás arrastrar o seleccionar los ficheros de tu ordenador que quieres subir. Pulsa en “browse” y se abrirá el cuadro para que selecciones los ficheros.

Navega hasta la carpeta donde descomprimiste los ficheros de datos y selecciona los tres ficheros CSV. Una vez que los selecciones, verás como comienza la carga.

Cuando haya terminado la carga, aparecerá una notificación bajo los ficheros

También verás dos botones. Si pulsas en “Create Table in Notebook”, Databricks creará automáticamente un nuevo cuaderno de trabajo con código de ejemplo de cómo leer estos ficheros desde un notebook usando PySpark. Hazlo y revisa el contenido.

1.4. USANDO CUADERNOS DE DATABRICKS.

Como ya hemos dicho, los cuadernos de trabajo de Databricks son muy similares a los de Jupyter. Hay celdas de texto, en las que escribir los comentarios, descripciones y anotaciones de lo que estamos haciendo, y celdas de código en las que podemos introducir las operaciones que queremos ejecutar.

En la parte superior del cuaderno encontrarás la barra del menú principal: muestra el nombre del cuaderno y varios submenús, incluyendo los de selección del cluster para ejecutar el cuaderno y el submenú ‘File’ para hacer copias del cuaderno, exportarlo o importar nuevo contenido desde fichero.

El primer elemento que aparece en la barra del menú a la izquierda es el cluster seleccionado actualmente para ejecutar el código del cuaderno. Cuando abres un cuaderno, por defecto verás que no hay seleccionado ningún cluster (aparece ‘Detached’).

Lo primero que debes hacer tras abrir el cuaderno, para poder trabajar con él, es conectarlo a un cluster. Abre la lista desplegable. Deberías ver el cluster que ya configuraste. Selecciónalo.

Si aparece marcado con un punto verde significa que el cluster ya está arrancado. Si no, en cuanto ejecutes la primera celda de código deberías ver como se arranca automáticamente (esto puede tardar varios segundos).

Sigamos con el cuaderno que se ha generado automáticamente al subir los ficheros de datos. La primera celda es un ejemplo de celda de texto. Por ahora no tienes que hacer nada con ellas, simplemente leer su contenido.

La segunda celda es una celda de código. En este caso, un ejemplo que muestra el código típico para leer los datos de uno de los ficheros que has subido y cargarlos en un dataframe de Spark.

La primera línea indica la ruta en la que Databricks ha almacenado el fichero. Fíjate bien en dicha ruta, porque todos los ficheros habrán sido almacenados en el mismo directorio (típicamente en "/FileStore/tables").

El resto de líneas definen distintas opciones para la operación de lectura del fichero en Spark. Estos comandos te deberían ser familiares.

```
# File location and type

file_location = "/FileStore/tables/prvdem.csv"

file_type = "csv"

# CSV options

infer_schema = "false"

first_row_is_header = "false"

delimiter = ","

# The applied options are for CSV files. For other file types, these will be ignored.

df = spark.read.format(file_type) \

    .option("inferSchema", infer_schema) \

    .option("header", first_row_is_header) \
```

```
.option("sep", delimiter) \

.load(file_location)display(df)
```

Para editar el código de una de estas celdas, simplemente haz click en su interior, verás cómo se activa el cursor de texto.

Para ejecutar el código de una celda tienes dos opciones. Hacer click en la opción “Run Cell”, en el botón de “Play” que encontrarás en la parte superior derecha de la celda:

O bien, habiendo entrado en la celda de código, puedes pulsar la combinación de teclas Control+Intro.

Prueba a ejecutar el código del ejemplo. Al completarse la ejecución de esta celda, verás el contenido del dataframe:

1.5. IMPORTAR EL CUADERNO DE TRABAJO DE LA SEGUNDA PARTE DEL EJERCICIO.

Para terminar esta primera parte y poder comenzar la siguiente, tienes que importar el fichero con el cuaderno de Databricks que contiene el enunciado del resto de tareas. Se trata del fichero

“Ejercicio Final - Big Data en la Nube - Cuaderno Enunciado.dbc”

Deberías encontrarlo con el resto de contenidos de este ejercicio final.

Una vez que tengas identificado el fichero en tu ordenador, vuelve a la pantalla principal de Databricks.

En la barra de navegación de la izquierda, selecciona la opción “Workspace”. Se desplegará un menú. En la opción “Shared” haz click sobre la flecha que aparece a la derecha. Verás varias opciones, entre ellas el comando “Import”. Selecciónalo.

Te aparecerá un nuevo diálogo para seleccionar la ubicación del fichero del cuaderno en tu ordenador.

Pulsa en “browse” y selecciona el archivo .dbc. Cuando aparezca una marca verde para indicar que se ha subido el fichero a la nube, pulsa en “Import”.

Una vez importado, puedes abrir el cuaderno desplegando el menú “Workspace” “Shared”. El cuaderno debería aparecerte en la carpeta compartida. Haz click en él para que se abra.

Las siguientes veces que quieras acceder, podrás encontrarlo en la pantalla principal nada más entrar (donde aparecen los ficheros recientes).

Ahora que ya tienes el cuaderno con el enunciado cargado, podemos continuar el ejercicio directamente en Databricks.