Data Science - Proyecto de Productivización

¡Bienvenidos al proyecto! Como sabréis os acaban de pedir una PoC (proof of concept) para un proyecto de Data Science. Al ser una PoC no se busca que esté perfecta, sino que mostréis el potencial del proyecto. Afortunadamente tenéis adjunto código de vuestro anterior proyecto en el que os podéis apoyar (está adjunto a este enunciado).

El departamento para el que es la PoC os ha dejado los datos en un repositorio (**podéis utilizar los datos que queráis**).

Nota: Vais a trabajar con PySpark, con un poco de comprensión de la nube donde está el cluster de PySpark y con alguna plataforma para alojar vuestro modelo. El modelado en sí no es clave y no es necesario obtener scorings excelentes.

Equipos

Trabajaréis toda la clase juntos y os dividiréis en 3 grupos. Los grupos empezarán a trabajar a la vez en paralelo.

GRUPO A: ETL con Spark

Debéis procesar los datos crudos (raw) con PySpark. Dichos datos crudos tendrán un campo de fecha y la precisión de los registros crudos serán segundos. Pueden existir tantos campos adicionales como deseéis. El formato inicial de los datos también es el que consideréis oportuno. Además de todas las fuentes de datos que ya conocéis, si necesitáis datos ficticios podéis utilizar el paquete de Python Faker.

Observaréis si esos datos crudos iniciales tienen sentido en vuestro problema, eliminando el ruido (valores sin sentido, outliers, nulos…). Generaréis tablas agregadas (por cuarto de hora, por hora…).

Aplicaréis feature engineering, cambio en el formato del fichero…

Antes de empezar esta parte es MUY RECOMENDABLE echar un ojo al código de vuestro proyecto anterior que viene adjunto a este documento.

Podéis utilizar Databricks Community Edition. Las capacidades que nos ofrece deberían ser suficientes.

GRUPO B: Modelado y análisis de la nube de Databricks

Por simplicidad y para ajustarnos a las limitaciones de Databricks CE el modelado se hará desde Pandas, SKlearn...
Echad un ojo a la parte de ML de Databricks, id en Databricks a la esquina superior izquierda y ahí seleccionad la parte de modelado, aunque vais a utilizar un modelado simple con DataFrames de Pandas, serializando el modelo. Pese a no tener los datos agregados limpios de partida, hablad con el GRUPO A para conocer el esquema y volumetría de los datos finales y empezad a modelar directamente. Luego tendréis que reentrenar el modelo con los datos reales.

Para el análisis de la nube, desde Databricks CE debéis analizar las características de la instancia del nodo maestro del cluster (donde tendréis el driver): hostname, en qué nube está, cuántos ejecutores tenemos en los nodos worker, qué usuario somos, en la instancia del nodo maestro cuál es la ruta hasta el driver, dónde está instalado Spark, qué base de datos relacional usa el metastore de Hive para saber dónde están ubicadas los datos que persiste en disco (bases de datos, tablas...), todo lo que encontréis analizando la instancia del nodo maestro.

GRUPO C: Creación de aplicación y subida a plataforma con nube

Desarrollaréis un frontend de una aplicación web basada en Flask. Hablaréis con el GRUPO B para saber qué modelo (qué entradas y qué salidas) vais a manejar. El GRUPO B os facilitará un modelo básico una

vez que ya tengáis diseñado el front. Incluiréis el modelo en la aplicación y lanzaréis la aplicación en local. Posteriormente subiréis la aplicación a PythonAnywhere, a Heroku o a otro servidor de aplicaciones web similar.

Plan de trabajo

El diseño de los grupos es para que avancéis todas las partes desde el principio, pero no todos los bloques tendrán la misma carga en los mismos momentos. Aquellos alumnos que finalicen su parte antes ayudarán en los otros grupos donde la carga de trabajo sea mayor.

Entregables

- PDF explicando el caso de uso

- Notebook de Databricks con las celdas comentadas y ejecutadas mostrando las ETL

- Documento PDF explicando qué información de la nube de Databricks habéis obtenido y cómo la habéis obtenido.

- Repositorio con la aplicación web Flask y PDF con el código e imágenes de ejecución de cada paso en el despliegue de la aplicación, así como la URL de acceso público.

¡A por todas! :-)