Proyecto1

Práctica de Spark

El objetivo de este taller es poner en práctica lo aprendido en el manejo de datos con Apache Spark. Para ello utilizaremos datos abiertos del estado colombiano que incluyen Contratos Electrónicos y Proyectos de Inversión. Tanto los datos originales como sus metadatos se encuentran en los siguientes enlaces:

- https://www.datos.gov.co/d/jbjy-vk9h
- https://www.datos.gov.co/d/cf9k-55fw

Tenga en cuenta que el campo BPIN identifica de forma única a los proyectos de inversión y por lo tanto es la llave que relaciona los dos conjuntos de datos.

Para acceder a los datos puede usar las siguientes ubicaciones:

- wasbs://sid@uniandesyjt.blob.core.windows.net/secop
- wasbs://sid@uniandesyjt.blob.core.windows.net/bpin

Tenga en cuenta que el conjunto de datos “Secop” se ha almacenado en la siguiente estructura (optimice sus consultas para que hagan uso de este particionamiento):

```
+---anno_firma=2015
|     *.json.gz
|
+---anno_firma=2016
|     *. json.gz
|
+---anno_firma=2017
|     *. json.gz
|
+--- ...
```

Mientras que BPIB tiene la siguiente estructura:

```
+---bpin
|     *.csv.gz
```

El taller debe ser entregado por **grupos** en **Bloque Neon,** para lo cual deben cargar un Notebook adecuadamente documentado que incluya los nombres de los estudiantes y resuelva uno a uno los siguientes puntos:

1) Lea los dos conjuntos de datos. Escriba el código para que la lectura sea lo más rápida posible y justifique su respuesta. **(9%)**

2) Identifique los 10 proveedores que han tenido el mayor valor de contratos durante el año 2024. **(13%)**

3) Identifique los 10 Proyectos de inversión que han tenido el mayor valor de contratos sin pagar (valor del contrato menos el valor pagado). La respuesta debe incluir el nombre de los proyectos y el valor sin pagar. **(13%)**

4) Identifique para cada año el top 5 de los proveedores con mayor valor de contratos. La respuesta debe incluir el nombre de los proveedores, el valor de sus contratos y su posición en el top. **(13%)**

5) Identifique el número de proveedores que firmaron contratos en el 2024 y NO firmaron contratos en el 2020. **(13%)**

6) Calcule el valor promedio de los contratos de "Prestación de servicios" y una aproximación de la mediana (el percentil 50). Construya esta respuesta de manera que se minimice el tiempo de cálculo, aunque se sacrifique precisión. **(13%)**

7) Cuál fue el año en el que firmaron contratos un mayor número de proveedores distintos. Construya esta respuesta de manera que se minimice el tiempo de cálculo, aunque se sacrifique precisión. **(13%)**

8) Cuáles son las 20 palabras más comunes en los objetos contractuales del 20% de los contratos más altos del 2020. Para esta respuesta omita stopwords (https://en.wikipedia.org/wiki/Stop_word, https://pypi.org/project/stop-words/), puntuaciones y no diferencie entre mayúsculas y minúsculas. **(13%)**

Recuerda que en la documentación encuentra información de funciones y parámetros para la carga de datos:

- Sobre la lectura de fuentes: https://spark.apache.org/docs/latest/sql-data-sources.html

- Sobre las funciones disponibles en pyspark: https://spark.apache.org/docs/latest/api/python/reference/pyspark.sql/functions.html

Si lo considere necesario utilice broadcast en los joins (https://spark.apache.org/docs/latest/api/python/reference/pyspark.sql/api/pyspark.sql.functions.broadcast.html).