Practica Nifi

https://localhost:8443/nifi

1) En el shell de Nifi, crear un script .sh que descargue el archivo titanic.csv al directorio /home/nifi/ingest (crearlo si es necesario). Ejecutarlo con `./home/nifi/ingest/ingest.sh`

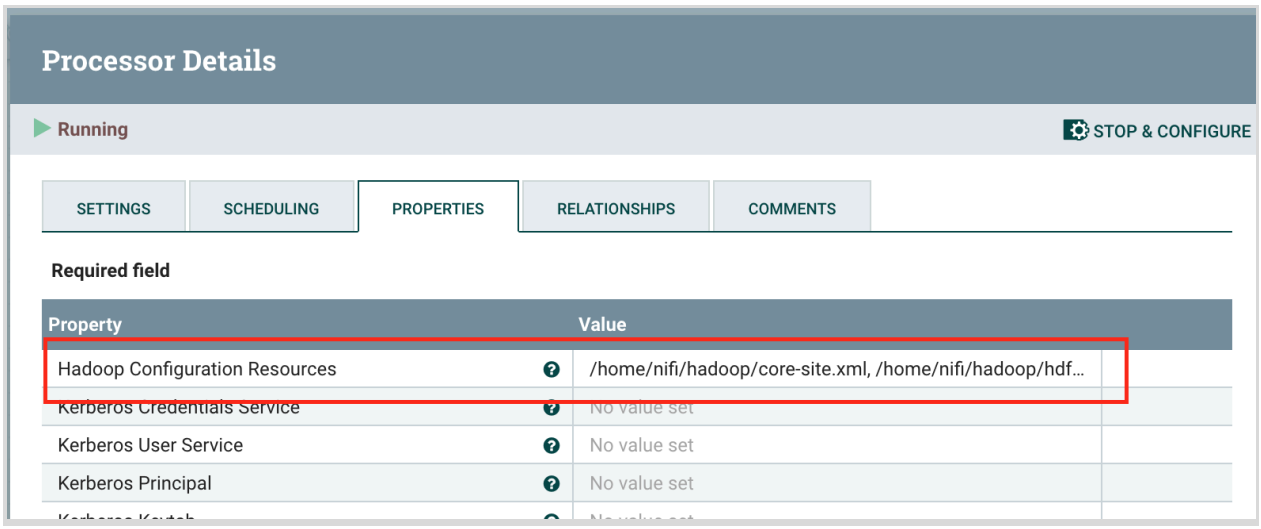
   https://dataengineerpublic.blob.core.windows.net/data-engineer/titanic.csv

2) Usando procesos en Nifi:
3) tomar el archivo titanic.csv desde el directorio /home/nifi/ingest.
4) Mover el archivo titanic.csv desde el directorio anterior, a /home/nifi/bucket (crear el directorio si es necesario)
5) Tomar nuevamente el archivo, ahora desde /home/nifi/bucket
6) Ingestarlo en HDFS/nifi (si es necesario, crear el directorio con `hdfs dfs -mkdir /nifi`)

Atencion:

- Para que Nifi pueda ingestar el archivo a HDFS, debe asignársele el permiso desde la consola de Hadoop con el comando `hdfs dfs -chmod 777 /nifi`
- Desde la consola de nifi, es necesario agregar dos archivos de configuración llamados core-site.xml y hdfs-site.xml al directorio /home/nifi/hadoop (crearlo si es necesario). Al final de este archivo está detallado cuál debe ser el contenido de ambos archivos
- Luego desde la interfaz gráfica de Nifi, al crear el proceso de ingesta de HDFS se debe definir en 'Properties/Hadoop Configuration Resources' la ruta a los archivos de configuración: `/home/nifi/hadoop/core-site.xml, /home/nifi/hadoop/hdfs-site.xml`

Diccionario de datos:
https://choens.github.io/titanic/workshops/regression/data-dictionary/



Core-site.xml de ejemplo

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
  Licensed under the Apache License, Version 2.0 (the "License");
  you may not use this file except in compliance with the License.
  You may obtain a copy of the License at

    http://www.apache.org/licenses/LICENSE-2.0

  Unless required by applicable law or agreed to in writing, software
  distributed under the License is distributed on an "AS IS" BASIS,
  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
  See the License for the specific language governing permissions and
  limitations under the License. See accompanying LICENSE file.
-->

<!-- Put site-specific property overrides in this file. -->


<configuration>
        <property>
                <name>fs.defaultFS</name>
                <value>hdfs://172.17.0.2:9000</value>
        </property>
</configuration>
```

Hdfs-site.xml de ejemplo

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
  Licensed under the Apache License, Version 2.0 (the "License");
  you may not use this file except in compliance with the License.
  You may obtain a copy of the License at

    http://www.apache.org/licenses/LICENSE-2.0

  Unless required by applicable law or agreed to in writing, software
  distributed under the License is distributed on an "AS IS" BASIS,
  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
  See the License for the specific language governing permissions and
  limitations under the License. See accompanying LICENSE file.
-->

<!-- Put site-specific property overrides in this file. -->


<configuration>

        <property>
                <name>dfs.replication</name>
                <value>1</value>
        </property>

        <property>
                <name>dfs.name.dir</name>
                <value>file:///home/hadoop/hadoopdata/hdfs/namenode</value>
        </property>

        <property>
                <name>dfs.data.dir</name>
                <value>file:///home/hadoop/hadoopdata/hdfs/datanode</value>
        </property>
</configuration>
```

7) Una vez que tengamos el archivo titanic.csv en HDFS realizar un pipeline en Airflow que ingeste este archivo y lo cargue en HIVE, teniendo en cuenta las siguientes transformaciones:
    a) Remover las columnas SibSp y Parch
    b) Por cada fila calcular el promedio de edad de los hombres en caso que sea hombre y promedio de edad de las mujeres en caso que sea mujer
    c) Si el valor de cabina en nulo, dejarlo en 0 (cero)

8) Una vez con la información en el datawarehouse calcular:
    a) Cuántos hombres y cuántas mujeres sobrevivieron
    b) Cuántas personas sobrevivieron según cada clase (Pclass)
    c) Cuál fue la persona de mayor edad que sobrevivió
    d) Cuál fue la persona más joven que sobrevivió