Practica Nifi

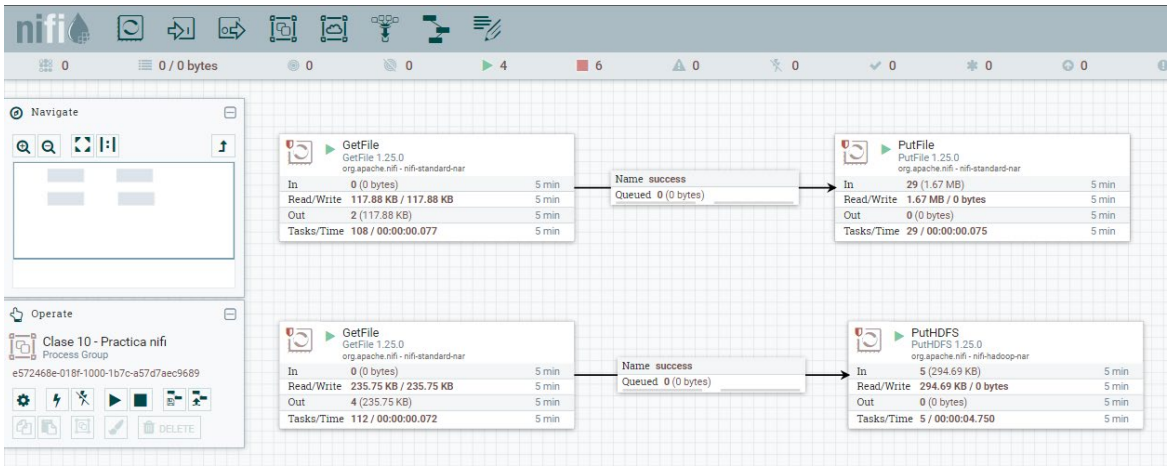
https://localhost:8443/nifi

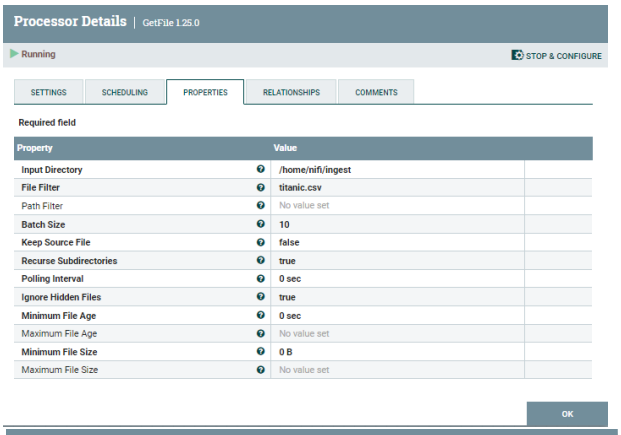
1) En el shell de Nifi, crear un script .sh que descargue el archivo titanic.csv al directorio /home/nifi/ingest (crearlo si es necesario). Ejecutarlo con ./home/nifi/ingest/ingest.sh

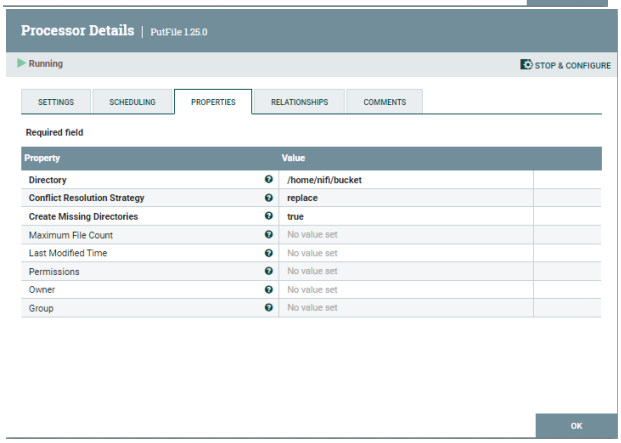
```
nifi@332d81692629:~/ingest$ ./ingest.sh
--2024-06-04 22:43:56--  https://dataengineerpublic.blob.core.windows.net/data-engineer/titanic.csv
Resolving dataengineerpublic.blob.core.windows.net (dataengineerpublic.blob.core.windows.net)... 20.150.25.164
Connecting to dataengineerpublic.blob.core.windows.net (dataengineerpublic.blob.core.windows.net)|20.150.25.164|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 60353 (59K) [text/csv]
Saving to: ‘/home/nifi/ingest/titanic.csv’

titanic.csv                         100%[=============================================================================>]  58.94K   358KB/s    in 0.2s

2024-06-04 22:43:58 (358 KB/s) - ‘/home/nifi/ingest/titanic.csv’ saved [60353/60353]

nifi@332d81692629:~/ingest$ cat ingest.sh
ruta='/home/nifi/ingest'

rm -f "${ruta}"titanic.csv

wget -P "${ruta}" https://dataengineerpublic.blob.core.windows.net/data-engineer/titanic.csv
nifi@332d81692629:~/ingest$
```

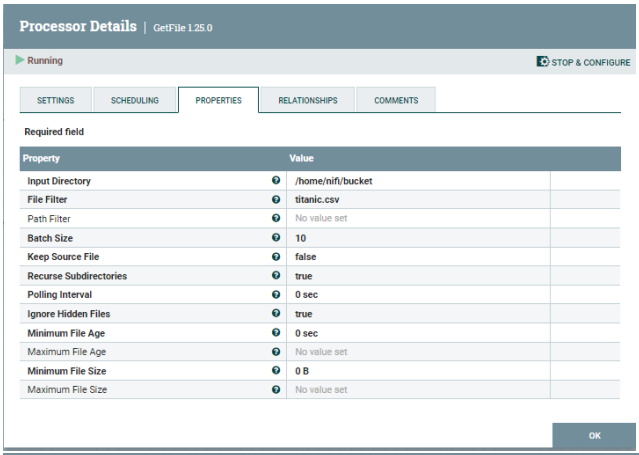
```
nifi@332d81692629:~/ingest$ ls
ingest_Parquet.sh  ingest.sh  titanic.csv
nifi@332d81692629:~/ingest$
```

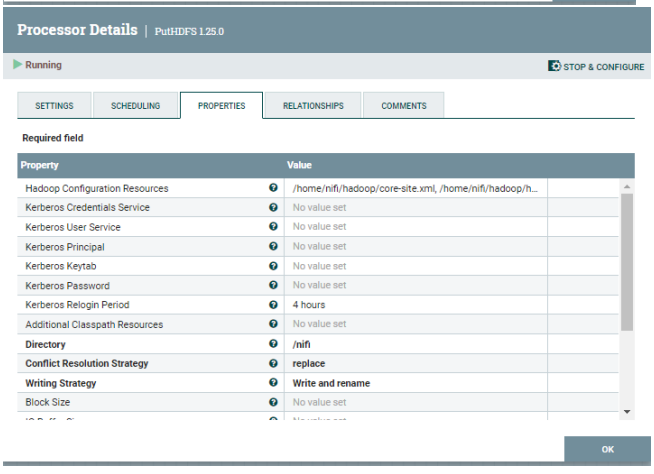
2) Usando procesos en Nifi:
3) tomar el archivo titanic.csv desde el directorio /home/nifi/ingest.
4) Mover el archivo titanic.csv desde el directorio anterior, a /home/nifi/bucket (crear el directorio si es necesario)
5) Tomar nuevamente el archivo, ahora desde /home/nifi/bucket
6) Ingestarlo en HDFS/nifi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




Processor Details | GetFile 1.25.0
Running
STOP & CONFIGURE
SETTINGS
SCHEDULING
PROPERTIES
RELATIONSHIPS
COMMENTS
Required field
Property
Value
Input Directory
/home/nifi/ingest
File Filter
titanic.csv
Path Filter
No value set
Batch Size
10
Keep Source File
false
Recurse Subdirectories
true
Polling Interval
0 sec
Ignore Hidden Files
true
Minimum File Age
0 sec
Maximum File Age
No value set
Minimum File Size
0 B
Maximum File Size
No value set
OK


Processor Details | PutFile 1.25.0
Running
STOP & CONFIGURE
SETTINGS
SCHEDULING
PROPERTIES
RELATIONSHIPS
COMMENTS
Required field
Property
Value
Directory
/home/nifi/bucket
Conflict Resolution Strategy
replace
Create Missing Directories
true
Maximum File Count
No value set
Last Modified Time
No value set
Permissions
No value set
Owner
No value set
Group
No value set
OK


Processor Details | GetFile 1.25.0
Running
STOP & CONFIGURE
SETTINGS
SCHEDULING
PROPERTIES
RELATIONSHIPS
COMMENTS
Required field
Property
Value
Input Directory
/home/nifi/bucket
File Filter
titanic.csv
Path Filter
No value set
Batch Size
10
Keep Source File
false
Recurse Subdirectories
true
Polling Interval
0 sec
Ignore Hidden Files
true
Minimum File Age
0 sec
Maximum File Age
No value set
Minimum File Size
0 B
Maximum File Size
No value set
OK


Processor Details | PutHDFS 1.25.0
Running
STOP & CONFIGURE
SETTINGS
SCHEDULING
PROPERTIES
RELATIONSHIPS
COMMENTS
Required field
Property
Value
Hadoop Configuration Resources
/home/nifi/hadoop/core-site.xml, /home/nifi/hadoop/h...
Kerberos Credentials Service
No value set
Kerberos User Service
No value set
Kerberos Principal
No value set
Kerberos Keytab
No value set
Kerberos Password
No value set
Kerberos Relogin Period
4 hours
Additional Classpath Resources
No value set
Directory
/nifi
Conflict Resolution Strategy
replace
Writing Strategy
Write and rename
Block Size
No value set
OK

```
hadoop@ec27db0d59e9:/$ hdfs dfs -ls /nifi
Found 1 items
-rw-r--r--   1 nifi supergroup      60353 2024-06-04 20:29 /nifi/titanic.csv
hadoop@ec27db0d59e9:/$
```

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

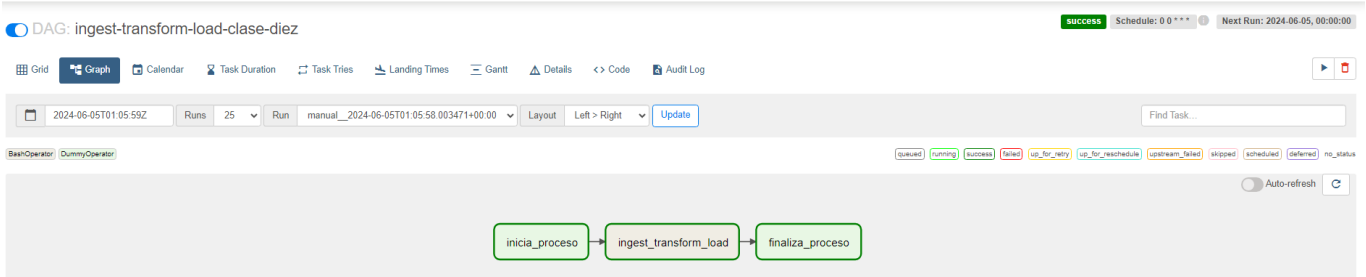
b) Por cada fila calcular el promedio de edad de los hombres en caso que sea hombre y promedio de edad de las mujeres en caso que sea mujer

c) Si el valor de cabina en nulo, dejarlo en 0 (cero)

```python
from pyspark.context import SparkContext
from pyspark.sql.session import SparkSession
from pyspark.sql import HiveContext
sc = SparkContext('local')
spark = SparkSession(sc)
hc = HiveContext(sc)

##leo csv de HDFS y lo cargo en un dataframe
df = spark.read.option("header", "true").csv("hdfs://172.17.0.2:9000/nifi/titanic.csv")

##creamos una vista del DF
df.createOrReplaceTempView("titanic")

##iltramos el DF para remover las columnas SibSp y Parch, calcular el promedio de edad y si el valor de cabina en nulo, dejarlo en 0 (cero)
df_titanic =  spark.sql("select cast(passengerId as int), cast(survived as int), cast(pclass as tinyint), cast(name as string), cast(sex as string), cast(age as int), cast(ticket as string), cast(fare as float), cast
(cabin as string), cast(embarked as string) from titanic")

df_titanic.createOrReplaceTempView("columnas_eliminadas")

df_titanic = spark.sql("select passengerId, survived, pclass, name, sex, age, ticket, fare, cabin, embarked, AVG(age) OVER (partition by sex) as avg_age_by_sex from columnas_eliminadas")

df_titanic.createOrReplaceTempView("columnas_eliminadas_promedio")

df_titanic = spark.sql("select passengerId, survived, pclass, CONCAT('\"', name, '\"') AS name, sex, age, ticket, fare, COALESCE(cabin, 0) as cabin, embarked, ROUND(avg_age_by_sex, 2) as avg_age_by_sex from
columnas_eliminadas_promedio ")

df_titanic.createOrReplaceTempView("vista_final_Load")

##insertamos el DF en la tabla titanic.information
spark.sql("insert into titanic.information select * from vista_final_Load")
```

CREATE EXTERNAL TABLE titanic.information(passengerId int, survived int, pclass tinyint, name string, sex string, age int, ticket string, fare float, cabin string, embarked string, avg_age_by_sex double)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
LOCATION '/tables/external/titanic/information';

```
hive> use titanic;
OK
Time taken: 0.04 seconds
hive> show tables;
OK
information
Time taken: 0.058 seconds, Fetched: 1 row(s)
hive>
```

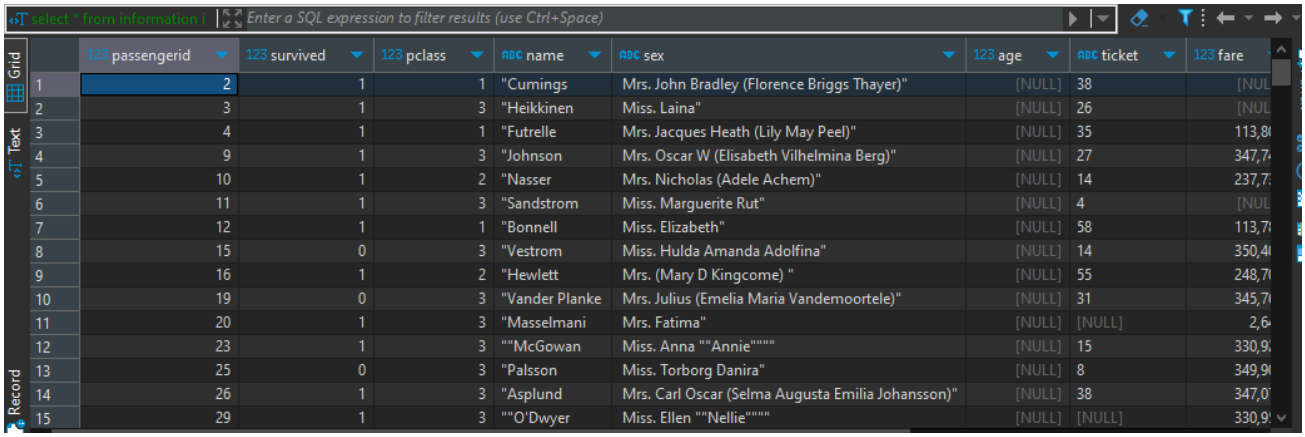


DAG: ingest-transform-load-clase-diez

Grid | Graph | Calendar | Task Duration | Task Tries | Landing Times | Gantt | Details | Code | Audit Log

```python
from datetime import timedelta
from airflow import DAG
from airflow.operators.bash import BashOperator
from airflow.operators.dummy import DummyOperator
from airflow.utils.dates import days_ago

args = {
    'owner': 'airflow',
}

with DAG(
    dag_id='ingest-transform-load-clase-diez',
    default_args=args,
    schedule_interval='0 0 * * *',
    start_date=days_ago(2),
    dagrun_timeout=timedelta(minutes=60),
    tags=['ingest', 'transform'],
    params={"example_key": "example_value"},
) as dag:

    inicia_proceso = DummyOperator(
        task_id='inicia_proceso',
    )

    finaliza_proceso = DummyOperator(
        task_id='finaliza_proceso',
    )

    ingest_transform_load = BashOperator(
        task_id='ingest_transform_load',
        bash_command='ssh hadoop@172.17.0.2 /home/hadoop/spark/bin/spark-submit --files /home/hadoop/hive/conf/hive-site.xml /home/hadoop/scripts/transform_clase_diez.py ',
    )


    inicia_proceso >> ingest_transform_load >> finaliza_proceso
```

Resultado en Hive:



select * from information_schema.tables | Enter a SQL expression to filter results (use Ctrl+Space)

Grid | Text | Record

	passengerid	survived	pclass	name	sex	age	ticket	fare
1	2	1	1	"Cumings	Mrs. John Bradley (Florence Briggs Thayer)"	[NULL]	38	[NUL
2	3	1	3	"Heikkinen	Miss. Laina"	[NULL]	26	[NUL
3	4	1	1	"Futrelle	Mrs. Jacques Heath (Lily May Peel)"	[NULL]	35	113,8
4	9	1	3	"Johnson	Mrs. Oscar W (Elisabeth Vilhelmina Berg)"	[NULL]	27	347,7
5	10	1	2	"Nasser	Mrs. Nicholas (Adele Achem)"	[NULL]	14	237,7
6	11	1	3	"Sandstrom	Miss. Marguerite Rut"	[NULL]	4	[NUL
7	12	1	1	"Bonnell	Miss. Elizabeth"	[NULL]	58	113,7
8	15	0	3	"Vestrom	Miss. Hulda Amanda Adolfina"	[NULL]	14	350,4
9	16	1	2	"Hewlett	Mrs. (Mary D Kingcome) "	[NULL]	55	248,7
10	19	0	3	"Vander Planke	Mrs. Julius (Emelia Maria Vandemoortele)"	[NULL]	31	345,7
11	20	1	3	"Masselmani	Mrs. Fatima"	[NULL]	[NULL]	2,6
12	23	1	3	""McGowan	Miss. Anna ""Annie"""	[NULL]	15	330,9
13	25	0	3	"Palsson	Miss. Torborg Danira"	[NULL]	8	349,9
14	26	1	3	"Asplund	Mrs. Carl Oscar (Selma Augusta Emilia Johansson)"	[NULL]	38	347,0
15	29	1	3	""O'Dwyer	Miss. Ellen ""Nellie"""	[NULL]	[NULL]	330,9