Practica Nifi

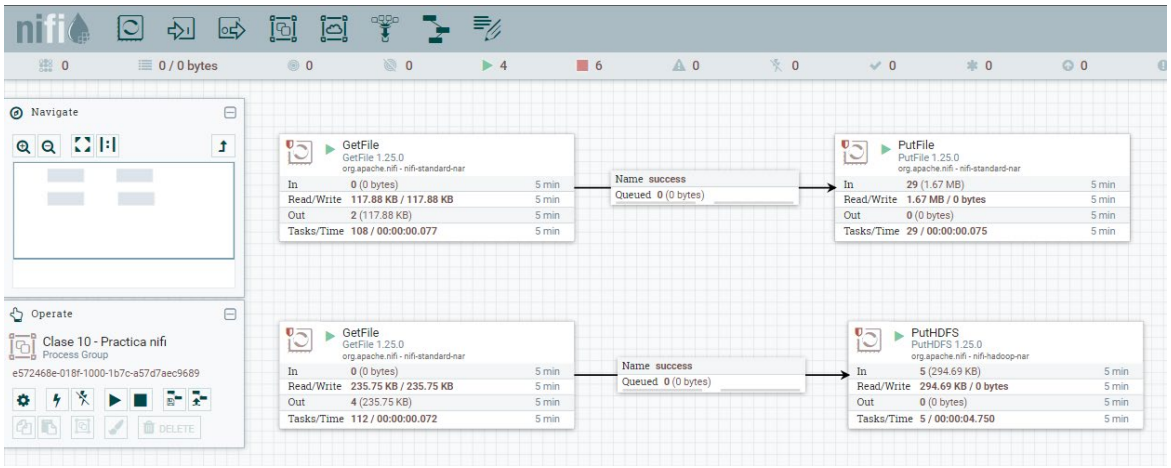
https://localhost:8443/nifi

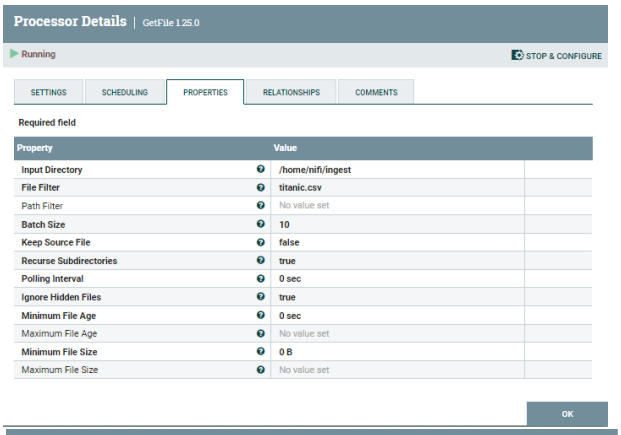
1) En el shell de Nifi, crear un script .sh que descargue el archivo titanic.csv al directorio /home/nifi/ingest (crearlo si es necesario). Ejecutarlo con ./home/nifi/ingest/ingest.sh

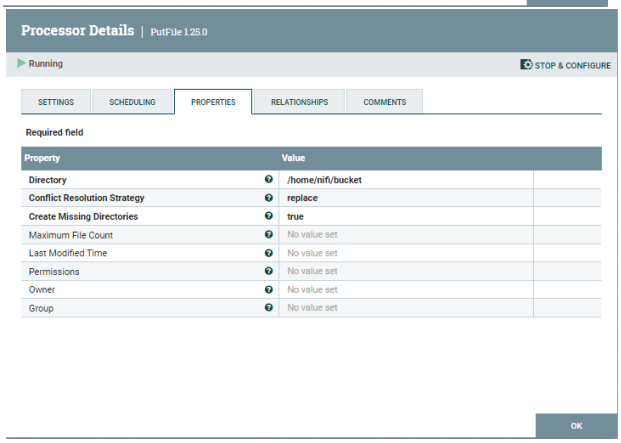
```
nifi@332d81692629:~/ingest$ ./ingest.sh
--2024-06-04 22:43:56--  https://dataengineerpublic.blob.core.windows.net/data-engineer/titanic.csv
Resolving dataengineerpublic.blob.core.windows.net (dataengineerpublic.blob.core.windows.net)... 20.150.25.164
Connecting to dataengineerpublic.blob.core.windows.net (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 "${ruta}" https://dataengineerpublic.blob.core.windows.net/data-engineer/titanic.csv
nifi@332d81692629:~/ingest$
```

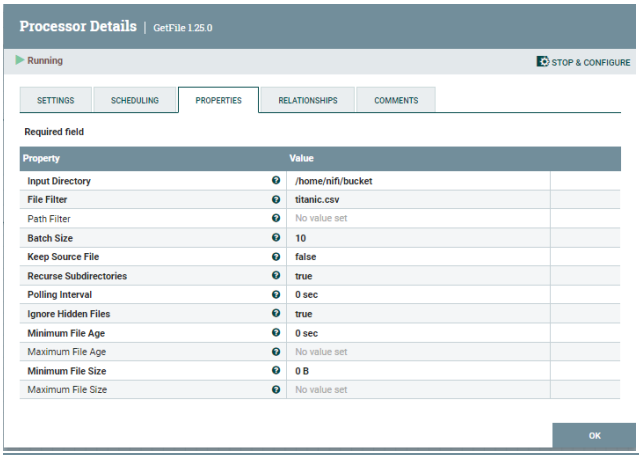
```
nifi@332d81692629:~/ingest$ ls
ingest_Parquet.sh  ingest.sh  titanic.csv
nifi@332d81692629:~/ingest$
```

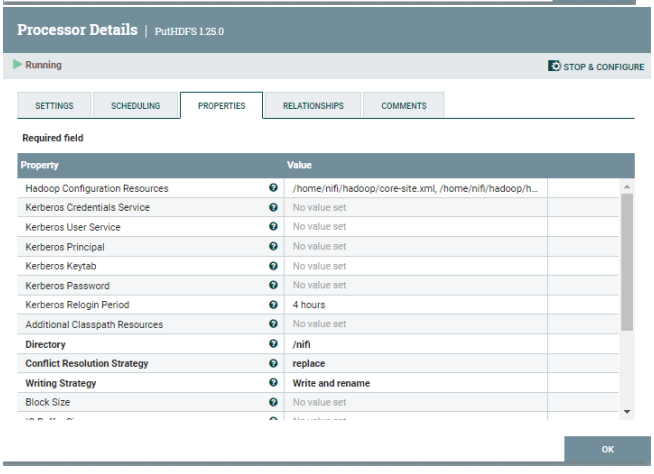
2) Usando procesos en Nifi:
3) tomar el archivo titanic.csv desde el directorio /home/nifi/ingest.
4) Mover el archivo titanic.csv desde el directorio anterior, a /home/nifi/bucket (crear el directorio si es necesario)
5) Tomar nuevamente el archivo, ahora desde /home/nifi/bucket
6) Ingestarlo en HDFS/nifi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

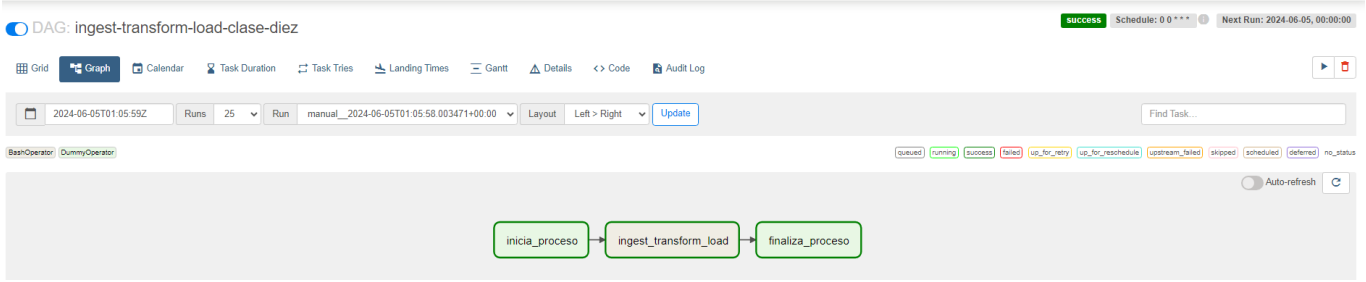
b) Por cada fila calcular el promedio de edad de los hombres en caso que sea hombre y promedio de edad de las mujeres en caso que sea mujer

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
from pyspark.sql.functions import regexp_replace, col

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 float), 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, 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 = df_titanic.withColumn('name', regexp_replace(col('name'), ',', ''))

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>
```

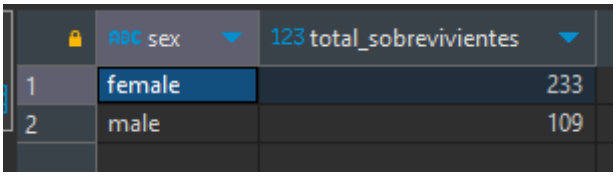



```python
from datetime import timedelta
from airflow import DAG
from airflow.operators.bash import BashOperator
from airflow.operators.dummy import DummyOperator
from airflow.utils.dates import days_ago

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
```

	passengerid	survived	pclass	name	sex	age	ticket	fare	cabin	embarked	avg_age_by_sex
4	494	0	1	Artagaveytia Mr. Ramon	male	71	PC 17609	49.5042	0	C	30.7
5	673	0	2	Mitchell Mr. Henry Michael	male	70	C.A. 24580	10.5	0	S	30.7
6	746	0	1	Crosby Capt. Edward Gifford	male	70	WE/P 5735	71	B22	S	30.7
7	117	0	3	Connors Mr. Patrick	male	70	370369	7.75	0	Q	30.7
8	34	0	2	Wheadon Mr. Edward H	male	66	C.A. 24579	10.5	0	S	30.7
9	457	0	1	Millet Mr. Francis Davis	male	65	13509	26.55	E38	S	30.7
10	55	0	1	Ostby Mr. Engelhart Cornelius	male	65	113509	61.9792	B30	C	30.7
11	281	0	3	Duane Mr. Frank	male	65	336439	7.75	0	Q	30.7
12	546	0	1	Nicholson Mr. Arthur Ernest	male	64	693	26	0	S	30.7
13	439	0	1	Fortune Mr. Mark	male	64	19950	263	C23 C25 C27	S	30.7
14	276	1	1	Andrews Miss. Kornelia Theodosia	female	63	13502	77.9583	D7	S	27.9
15	484	1	3	Turkula Mrs. (Hedwig)	female	63	4134	9.5875	0	S	27.9
16	253	0	1	Stead Mr. William Thomas	male	62	113514	26.55	C87	S	30.7
17	830	1	1	Stone Mrs. George Nelson (Martha Evelyn)	female	62	113572	80	B28	[NULL]	27.9
18	571	1	2	Harris Mr. George	male	62	S.W./PP 752	10.5	0	S	30.7
19	556	0	1	Wright Mr. George	male	62	113807	26.55	0	S	30.7
20	626	0	1	Sutton Mr. Frederick	male	61	36963	32.3208	D50	S	30.7
21	171	0	1	Van der hoef Mr. Wyckoff	male	61	111240	33.5	B19	S	30.7
22	327	0	3	Nysveen Mr. Johan Hansen	male	61	345364	6.2375	0	S	30.7
23	695	0	1	Weir Col. John	male	60	113800	26.55	0	S	30.7
24	685	0	2	Brown Mr. Thomas William Solomon	male	60	29750	39	0	S	30.7
25	367	1	1	Warren Mrs. Frank Manley (Anna Sophia Atl	female	60	110813	75.25	D37	C	27.9
26	588	1	1	Frolicher-Stehli Mr. Maxmillian	male	60	13567	79.2	B41	C	30.7
27	233	0	2	Sjostedt Mr. Ernst Adolf	male	59	237442	13.5	0	S	30.7
28	95	0	3	Coxon Mr. Daniel	male	59	364500	7.25	0	S	30.7
29	488	0	1	Kent Mr. Edward Austin	male	58	11771	29.7	B37	C	30.7
30	660	0	1	Newell Mr. Arthur Webster	male	58	35273	113.275	D48	C	30.7
31	12	1	1	Bonnell Miss. Elizabeth	female	58	113783	26.55	C103	S	27.9
32	269	1	1	Graham Mrs. William Thompson (Edith Junk	female	58	PC 17582	153.4625	C125	S	27.9
33	196	1	1	Lurette Miss. Elise	female	58	PC 17569	146.5208	B80	C	27.9
34	773	0	2	Mack Mrs. (Mary)	female	57	S.O./P.P. 3	10.5	E77	S	27.9
35	627	0	2	Kirkland Rev. Charles Leonard	male	57	219533	12.35	0	Q	30.7
36	880	1	1	Potter Mrs. Thomas Jr (Lily Alexenia Wilson)	female	56	11767	83.1583	C50	C	27.9
37	468	0	1	Smart Mr. John Montgomery	male	56	113792	26.55	0	S	30.7

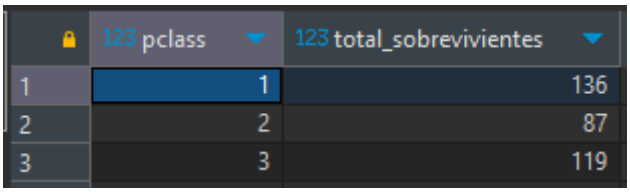
8) Una vez con la información en el datawarehouse calcular:

a) Cuántos hombres y cuántas mujeres sobrevivieron



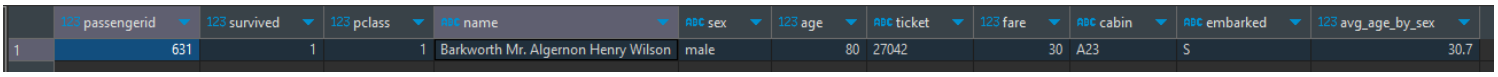
	sex	total_sobrevivientes
1	female	233
2	male	109

b) Cuántas personas sobrevivieron según cada clase (Pclass)



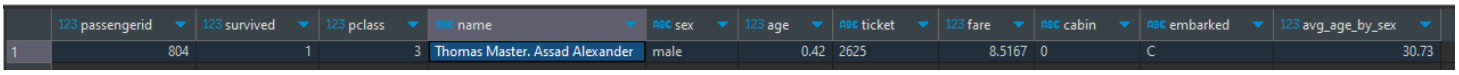
	pclass	total_sobrevivientes
1	1	136
2	2	87
3	3	119

c) Cuál fue la persona de mayor edad que sobrevivió



	passengerid	survived	pclass	name	sex	age	ticket	fare	cabin	embarked	avg_age_by_sex
1	631	1	1	Barkworth Mr. Algernon Henry Wilson	male	80	27042	30	A23	S	30.7

d) Cuál fue la persona más joven que sobrevivió



	passengerid	survived	pclass	name	sex	age	ticket	fare	cabin	embarked	avg_age_by_sex
1	804	1	3	Thomas Master. Assad Alexander	male	0.42	2625	8.5167	0	C	30.73