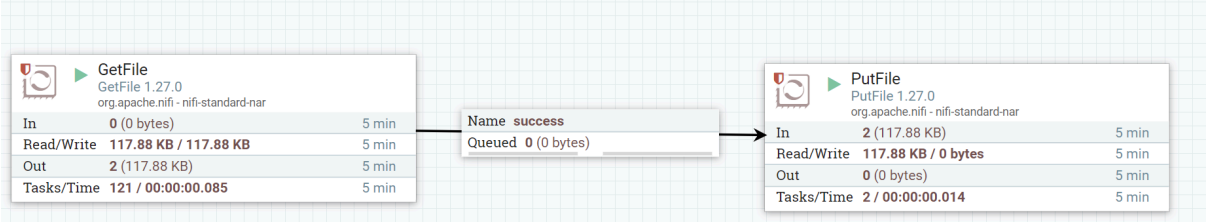
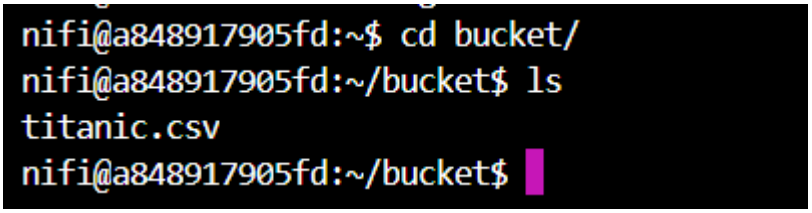
1.

```
nifi@a848917905fd:~/ingest$ ls
ingest.sh
nifi@a848917905fd:~/ingest$ cat ingest.sh
#!/bin/bash

# Directorio de destino
DEST_DIR="/home/nifi/ingest"

# Crear el directorio si no existe
if [ ! -d "$DEST_DIR" ]; then
    echo "Creando el directorio $DEST_DIR"
    mkdir -p "$DEST_DIR"
fi

# Descargar el archivo titanic.csv
FILE_URL="https://dataengineerpublic.blob.core.windows.net/data-engineer/titanic.c
DEST_FILE="$DEST_DIR/titanic.csv"

echo "Descargando el archivo titanic.csv a $DEST_FILE"
curl -o "$DEST_FILE" "$FILE_URL"

# Verificar si la descarga fue exitosa
if [ $? -eq 0 ]; then
    echo "Descarga completada exitosamente."
else
    echo "Error al descargar el archivo."
    exit 1
fi
nifi@a848917905fd:~/ingest$
```

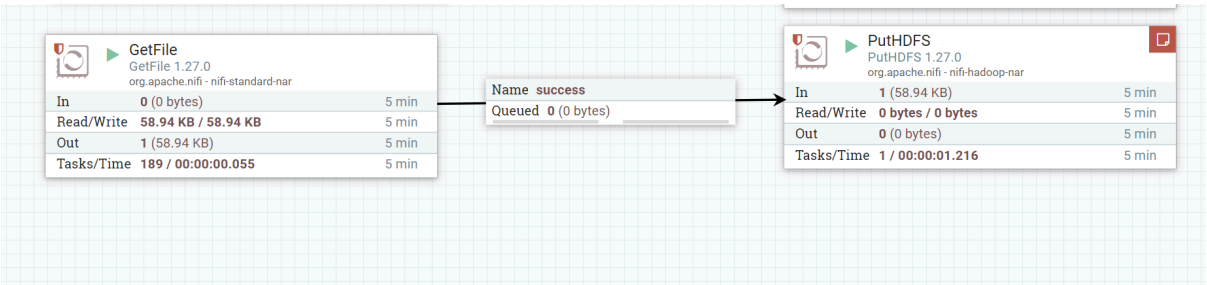
```
nifi@a848917905fd:~/ingest$ ./ingest.sh
Descargando el archivo titanic.csv a /home/nifi/ingest/titanic.csv
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100 60353  100 60353    0     0  67225      0 --:--:-- --:--:-- --:--:-- 67208
Descarga completada exitosamente.
nifi@a848917905fd:~/ingest$ ls
ingest.sh  titanic.csv
nifi@a848917905fd:~/ingest$
```

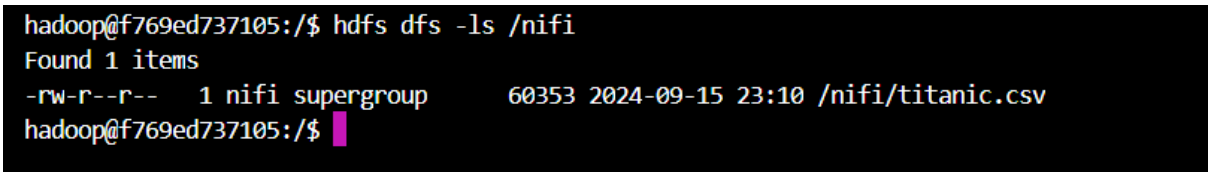
2, 3 y 4


GetFile
GetFile 1.27.0
org.apache.nifi - nifi-standard-nar
In 0 (0 bytes) 5 min
Read/Write 117.88 KB / 117.88 KB 5 min
Out 2 (117.88 KB) 5 min
Tasks/Time 121 / 00:00:00.085 5 min
Name success
Queued 0 (0 bytes)
PutFile
PutFile 1.27.0
org.apache.nifi - nifi-standard-nar
In 2 (117.88 KB) 5 min
Read/Write 117.88 KB / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 2 / 00:00:00.014 5 min


nifi@a848917905fd:~$ cd bucket/
nifi@a848917905fd:~/bucket$ ls
titanic.csv
nifi@a848917905fd:~/bucket$

5 y 6

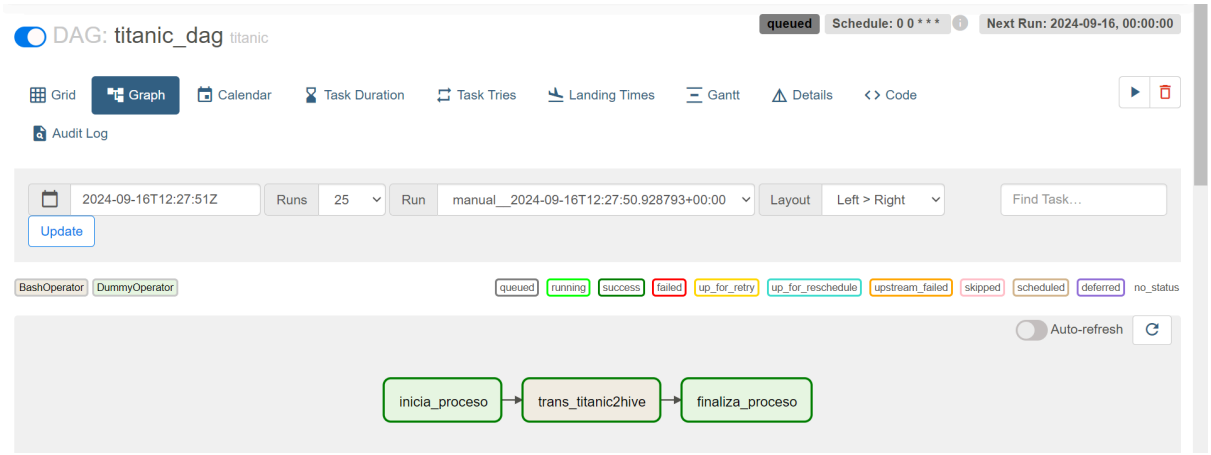

GetFile
GetFile 1.27.0
org.apache.nifi - nifi-standard-nar
In 0 (0 bytes) 5 min
Read/Write 58.94 KB / 58.94 KB 5 min
Out 1 (58.94 KB) 5 min
Tasks/Time 189 / 00:00:00.055 5 min
Name success
Queued 0 (0 bytes)
PutHDFS
PutHDFS 1.27.0
org.apache.nifi - nifi-hadoop-nar
In 1 (58.94 KB) 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 1 / 00:00:01.216 5 min


hadoop@f769ed737105:/$ hdfs dfs -ls /nifi
Found 1 items
-rw-r--r-- 1 nifi supergroup 60353 2024-09-15 23:10 /nifi/titanic.csv
hadoop@f769ed737105:/$

7

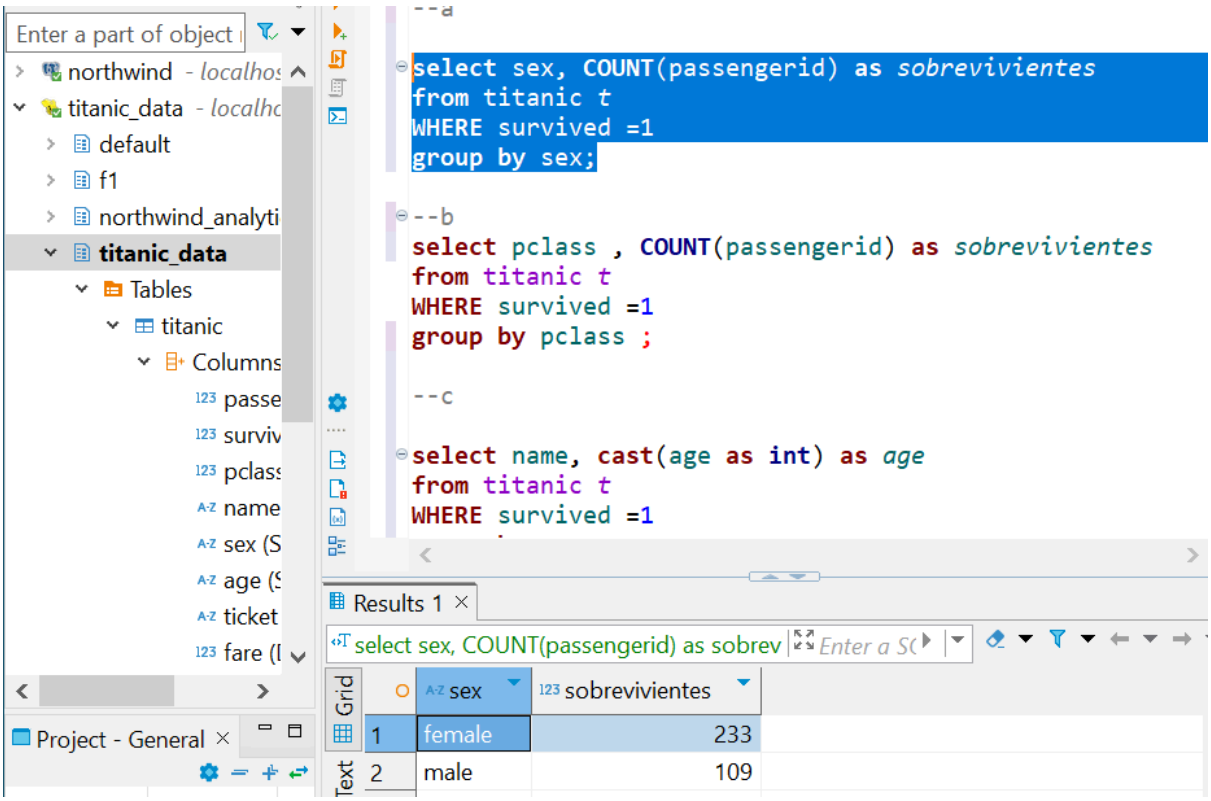
```
hadoop@f769ed737105:~/scripts$ cat trans_titanic.py
from pyspark.context import SparkContext
from pyspark.sql.session import SparkSession
from pyspark.sql.functions import col, avg, when
sc = SparkContext('local')
spark = SparkSession(sc)
from pyspark.sql import HiveContext
hc = HiveContext(sc)


titanic = spark.read.csv('hdfs://172.17.0.2:9000/nifi/titanic.csv', header=True, inferSchema=True)


titanic = titanic.drop('SibSp', 'Parch')


avg_ages = titanic.groupBy('Sex').agg(avg('Age').alias('avg_age')).collect()
avg_male_age = [row['avg_age'] for row in avg_ages if row['Sex'] == 'male'][0]
avg_female_age = [row['avg_age'] for row in avg_ages if row['Sex'] == 'female'][0]

titanic = titanic.withColumn('Age', when((col('Sex') == 'male') & col('Age').isNull(), avg_male_age)
                            .when((col('Sex') == 'female') & col('Age').isNull(), avg_female_age)
                            .otherwise(col('Age')))


titanic = titanic.fillna({'Cabin': 0})


titanic.write.insertInto('titanic_data.titanic')

spark.stop()
```

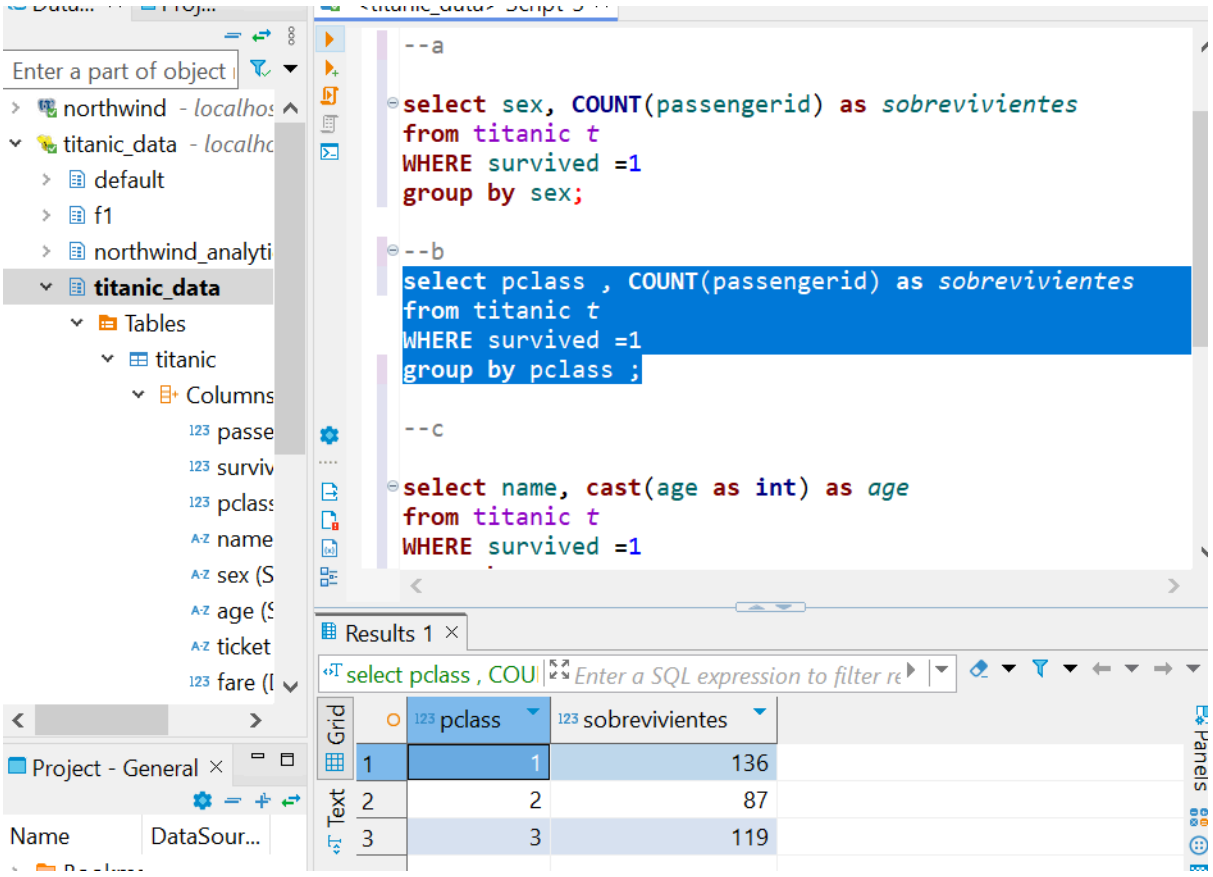
```
from datetime import timedelta
from airflow import DAG
from airflow.operators.bash_operator import BashOperator
from airflow.operators.dummy import DummyOperator
from airflow.utils.dates import days_ago


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

with DAG(
    dag_id='titanic_dag',
    default_args= args,
    description='titanic',
    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',
    )

    trans_titanic2hive = BashOperator(
            task_id='trans_titanic2hive',
            bash_command='ssh hadoop@172.17.0.2 /home/hadoop/spark/bin/spark-submit --files /home/ha
doop/hive/conf/hive-site.xml /home/hadoop/scripts/trans_titanic.py ',
        )

    finaliza_proceso = DummyOperator(
        task_id='finaliza_proceso',
    )
```

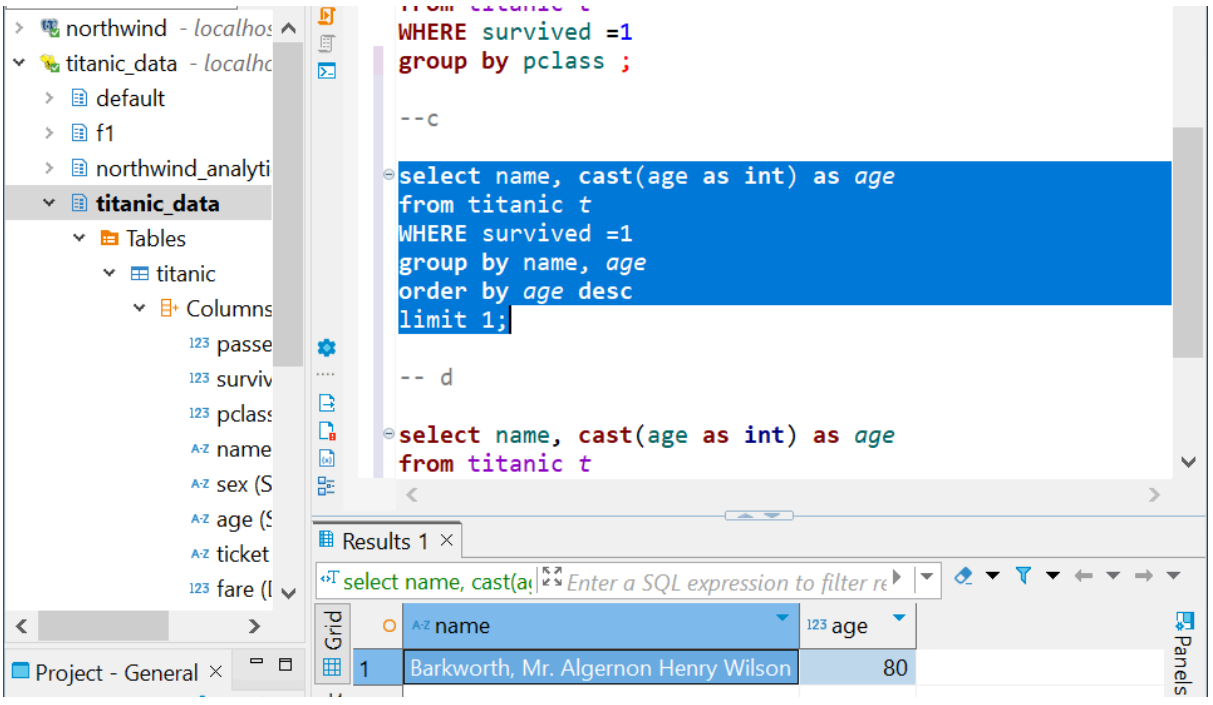


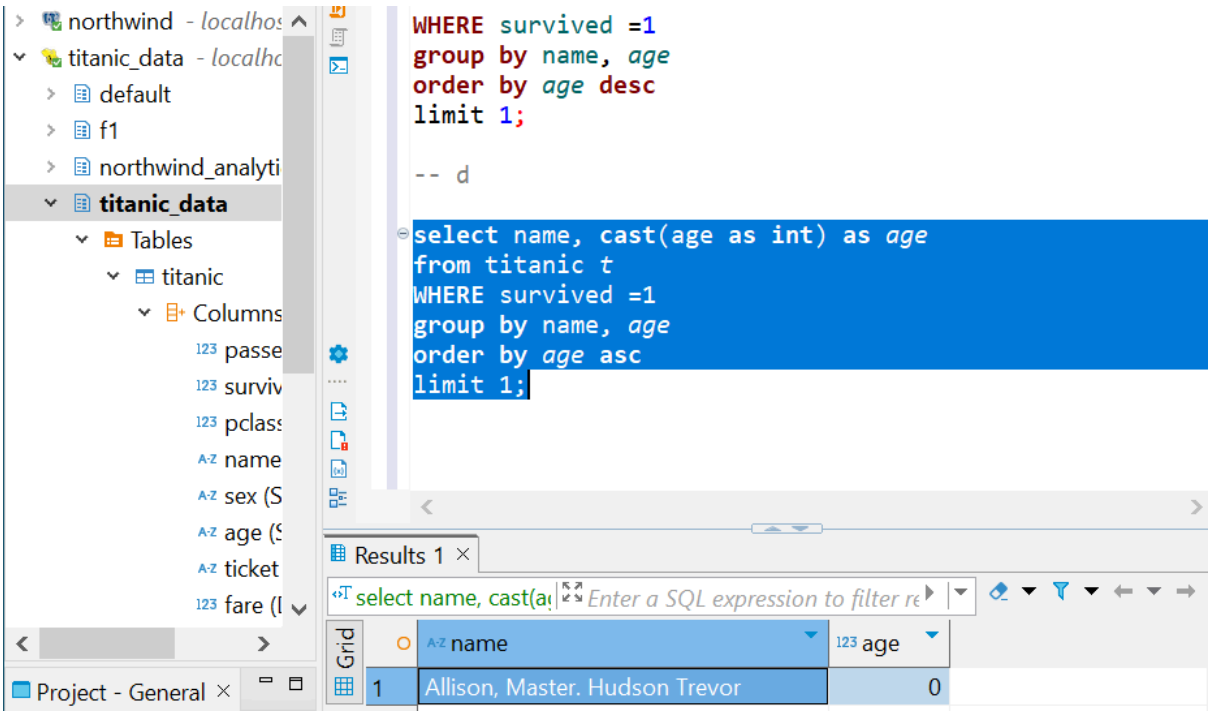

```
hive> select * from titanic limit 5;
OK
1       0       3       Braund   Mr. Owen Harris          male    22      NULL    7.25            0
2       1       1       Cumings  Mrs. John Bradley (Florence Briggs Thayer)    female          3
8       NULL    71.2833 C85
3       1       3       Heikkinen        Miss. Laina     female  26      NULL    7.925           0
4       1       1       Futrelle         Mrs. Jacques Heath (Lily May Peel)    female          3
5       113803.0        53.1    C123
5       0       3       Allen    Mr. William Henry       male    35      373450.0                8
.05     0
Time taken: 0.19 seconds, Fetched: 5 row(s)
```

8




Enter a part of object
northwind - localhos
titanic_data - localhc
default
f1
northwind_analyti
titanic_data
Tables
titanic
Columns
passe
surviv
pclass
name
sex (S
age (S
ticket
fare (I
Project - General
Name
DataSour...
--a
select sex, COUNT(passengerid) as sobrevivientes
from titanic t
WHERE survived =1
group by sex;
--b
select pclass , COUNT(passengerid) as sobrevivientes
from titanic t
WHERE survived =1
group by pclass ;
--c
select name, cast(age as int) as age
from titanic t
WHERE survived =1
Results 1
select pclass , COU
Enter a SQL expression to filter re
Grid
Text
Panels
pclass
sobrevivientes
1 1 136
2 2 87
3 3 119


northwind - localhos
titanic_data - localhc
default
f1
northwind_analyti
titanic_data
Tables
titanic
Columns
passe
surviv
pclass
name
sex (S
age (S
ticket
fare (I
Project - General
WHERE survived =1
group by pclass ;
--c
select name, cast(age as int) as age
from titanic t
WHERE survived =1
group by name, age
order by age desc
limit 1;
-- d
select name, cast(age as int) as age
from titanic t
Results 1
select name, cast(a
Enter a SQL expression to filter re
Grid
Panels
name
age
1 Barkworth, Mr. Algernon Henry Wilson 80


northwind - localhos
titanic_data - localhc
default
f1
northwind_analyti
titanic_data
Tables
titanic
Columns
passe
surviv
pclass
name
sex (S
age (S
ticket
fare (I
Project - General
WHERE survived =1
group by name, age
order by age desc
limit 1;
-- d
select name, cast(age as int) as age
from titanic t
WHERE survived =1
group by name, age
order by age asc
limit 1;
Results 1
select name, cast(a
Enter a SQL expression to filter re
Grid
name
age
1
Allison, Master. Hudson Trevor
0