Extract, Transform, Load

Team Members:

Albert Aninagyei Ofori, Bruna Goncalves, Dameon Turner,

Ralph Watson-Quartey, Geo Mirador

Project Description

Create: An actor's table with unique actors and actress' names that also contains a numerical field for age and a boolean field to identify if the actor or actress is currently alive, deceased, or unknown.

Create: A film table that contains a character field for the original title, a numerical field for the average critic rating, a character field for the type of film (movie, television show, etc.) and a datetime field that has the runtime of the film in HH:MM (hours then minutes, separated by a colon) format. In addition to the required fields, there should be additional fields for other relevant metadata.

Create: A producer's table that contains only producers names and links to the film table.

Create: A characters table with unique character names that links the actors and film tables together.

Actors Table

```
spark.sql('select nconst, primaryname, status, age  from actors_final').show()
```

nconst	primaryname	status	age
nm0462116	Jeff Kober	1	68.0
nm0462319	Herman Koch	1	68.0
nm7741448	Thomas Ganidel	null	null
nm7744565	Charles-André Lac...	null	null
nm7752147	Gaihlin St-Onge	null	null
nm1443327	Robert Desrosiers	null	null
nm1443573	Alan Stone	null	null
nm1445190	Siu-Yan Cha	null	null
nm1445529	Glenn Hanning	null	null
nm1443396	Fabian Hinrichs	1	47.0
nm1445313	Kate Dorrington	null	null
nm1445861	Christopher Marti...	0	68.0
nm4943457	Pierson Fode	1	30.0
nm4946463	Derek Yates	null	null
nm7743441	Jordan Kent	null	null
nm10586905	Sheena Catacutan	null	null
nm10589858	Crystal Cook	null	null
nm3126747	Aayam Mehta	null	null
nm3130303	Aj Platt	null	null
nm3130906	Little Don	null	null

only showing top 20 rows

Create an actor's table with unique actors and actress' names that also contains a numerical field for age and a boolean field to identify if the actor or actress is currently alive, deceased, or unknown.

Alive/Dead/Unknown

```
# Create table of all actors
# If birth year is known and death year is known, actor is dead,(0)
# If birth year is known and death year is unknown, then actor is assumed to be alive,(1)
# Otherwise actor status is null(WDN)
```

Age of actors

```
WHEN STATUS = 0 (dead) and birthyear IS NOT NULL THEN deathyear - birthyear

WHEN STATUS = 1 (Alive)THEN 2021 - birthyear
ELSE NULL (WDN)END AS AGE from actors
```

Film Table

```
[22] # Left join all movies in the titles table with their corresponding ratings from the ratings data.
     titles_final = spark.sql('select titles.tconst, titles.originaltitle, ratings.averagerating, titles.ti
     titles_final.show()
```

```
+---------+--------------------+-------------+---------+-------+-------+
|   tconst|       originaltitle|averagerating|titletype|isadult|runtime|
+---------+--------------------+-------------+---------+-------+-------+
|tt0439997|           500 Almas|          7.2|    movie|      0| 001:45|
|tt0439999|           80s Mania|         null|tvSpecial|      0| 000:50|
|tt0440003|         A1 tou tiao|          5.9|    movie|      0| 001:35|
|tt0440004| AD/BC: A Rock Opera|          7.4|  tvMovie|      0| 000:30|
|tt0440008|Abbamania: We Say...|         null|  tvMovie|      0| 000:50|
|tt0440016|       Ah ma yau nan|          5.5|    movie|      0| 001:33|
|tt0440022|        Al atardecer|         null|  tvMovie|      0| 001:06|
|tt0440035|      L'amour en pen|         null|  tvMovie|      0| 000:52|
|tt0440067|      Bau lit do see|          5.4|    movie|      0| 001:39|
|tt0440078|  The Band Aid Story|          8.1|  tvMovie|      0| 001:35|
|tt0440084|A Beachcombers Ch...|          7.0|  tvMovie|      0| 002:00|
|tt0440149|Blink 182: Punk P...|          4.1|    video|      0| 001:01|
|tt0440154|Boogie special: 5...|          1.0|  tvMovie|      0| 000:26|
|tt0440155|Boogie special: M...|          7.7|  tvMovie|      0| 000:28|
|tt0440157|The British Comed...|          6.2|tvSpecial|      0| 002:05|
|tt0440158|      Broken Bridges|          7.9|    movie|      0| 001:36|
|tt0440166|      Bump and Grind|         null|  tvMovie|      0| 000:46|
|tt0440167|   Burgers/Reizigers|          6.3|  tvMovie|      0| 001:00|
|tt0440174|CD Hoy: Portraits...|         null|    movie|      0| 000:48|
|tt0440175|          Cabra-Cega|          6.3|    movie|      0| 001:47|
+---------+--------------------+-------------+---------+-------+-------+
only showing top 20 rows
```

Create a film table that contains a character field for the original title,

Create a numerical field for the average critic rating,

Create a character field for the type of film and a datetime field that has the runtime of the film in HH:MM (hours then minutes, separated by a colon) format.

HH:MM Calculation

```
concat(RIGHT(concat("000",cast(titles.runtimeminutes/60 as int)),3),":",RIGHT(concat("00",cast(titles.runtimeminutes%60 as int)),2)) as runtime from titles left join ratings on titles.tconst = ratings.tconst'
```

Producer Table

```
[28] producers_final.show()
     producers_final.createOrReplaceTempView('prod_final')
```

```
+---------+--------------------+----------+
|   nconst|         primaryname|    tconst|
+---------+--------------------+----------+
|nm7745824|         Poppy Begum| tt6400730|
|nm7745824|         Poppy Begum| tt7610596|
|nm7745824|         Poppy Begum| tt4116046|
|nm7745824|         Poppy Begum|tt10196182|
|nm7748384|        Woody Daigle| tt4687782|
|nm1447145|Jennifer M. Fah-V...| tt0896084|
|nm1447145|Jennifer M. Fah-V...| tt0961102|
|nm1447145|Jennifer M. Fah-V...| tt0437758|
|nm1443310|      Michael Cotter| tt2177489|
|nm1443310|      Michael Cotter| tt1791528|
|nm1443310|      Michael Cotter| tt0285403|
|nm1443310|      Michael Cotter| tt0805666|
|nm1447108|        Dan Mogensen| tt1042453|
|nm1447108|        Dan Mogensen| tt2126045|
|nm1447108|        Dan Mogensen| tt5896744|
|nm1447108|        Dan Mogensen| tt3281048|
|nm4943760|         Tom Gretzer| tt2294661|
|nm4947076|      Kerstin Freels| tt9262068|
|nm4947076|      Kerstin Freels|tt11701822|
|nm1447447|      Brian Schulman| tt6217804|
+---------+--------------------+----------+
only showing top 20 rows
```

Create: A producer's table that contains only producers names and links to the film table.

```
# filterNames where primaryprofession like '%producer%'
# Select all producers and the foreign key to every title for which they are known
```

Character Table

```
# Display the characters and their corresponding movies and actors
characters_final = characters.dropDuplicates()
characters_final.show()
```

```
+--------------------+----------+----------+
|          characters|    tconst|    nconst|
+--------------------+----------+----------+
| ["""Pomegranate"""]"| tt1725077| nm4086624|
|      """Fraile"""]"| tt0446914| nm0781103|
|     """Penalty"""]"| tt8975184| nm3067739|
|"["""Self - Contes...|tt13984052|nm12242252|
| ["""Bimbo Coles"""]"| tt0451016| nm1871139|
|        """Self"""]"| tt0443374| nm2092808|
|        """Self"""]"|tt12252890| nm1432434|
|       """Angel"""]"| tt2034049| nm0000327|
|         """Zac"""]"| tt6072502| nm8872970|
|   """Lady Gaga"""]"| tt6743882| nm3078932|
|"["""Shalini (Boss...| tt9547758| nm1427076|
|  """Candy Kiss"""]"|tt10743654| nm6629894|
|"["""Duke Sagrado"""]"| tt1781844| nm0320282|
|"["""Martin Solvei...|tt10479256| nm1528519|
|"["""Self - Dancer...| tt2048152| nm2612563|
|"["""Robert Braula...| tt6607896| nm8825458|
| """Pvt. Cooper"""]"| tt0280609| nm0571727|
|   """Young boy"""]"| tt8582012| nm7640138|
|        """Self"""]"| tt1797546| nm0776441|
|        """Self"""]"|tt11559474| nm1935086|
+--------------------+----------+----------+
only showing top 20 rows
```

Create: A characters table with unique character names that links the actors and film tables together.

```
# select characters, tconst, nconst from principals where category in ("actor", "actress", "self") and characters IS NOT NULL
```

Entity Relationship Diagram

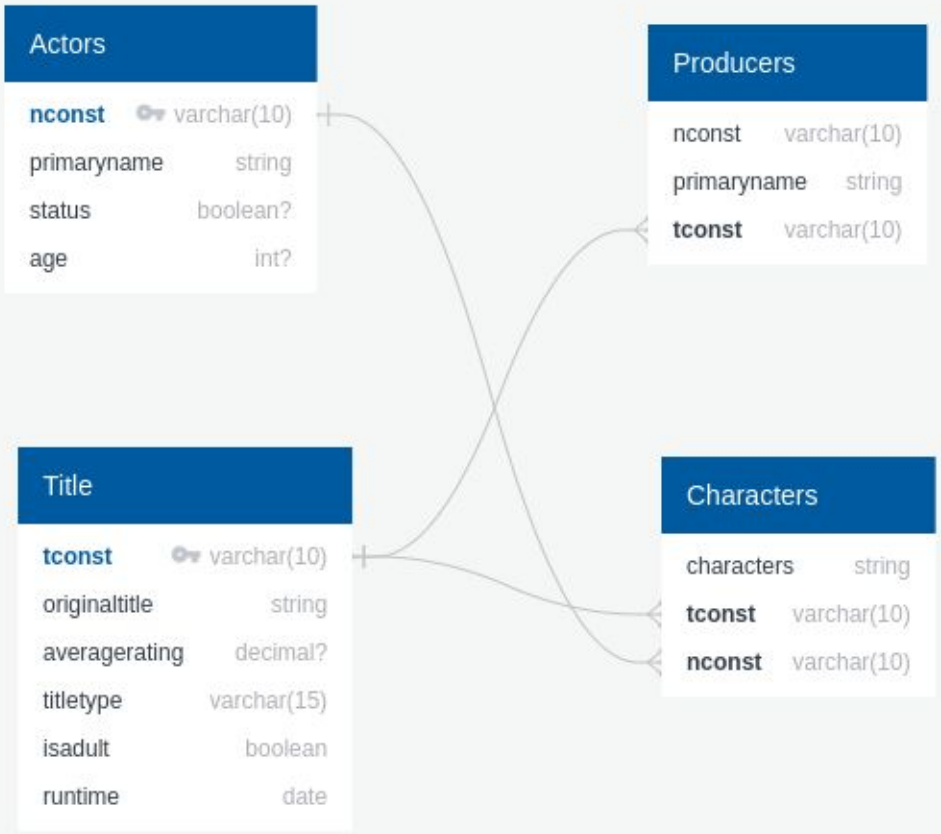

TABLE SCHEMAS

Actors
-
nconst PK varchar(10)
primaryname string
status NULL boolean
age NULL int

Title
-
tconst PK varchar(10)
originaltitle string
averagerating NULL decimal
titletype varchar(15)
isadult boolean
runtime date

Producers
---
nconst varchar(10)
primaryname string
tconst varchar(10) FK >- Title.tconst

Characters
------
characters string
tconst varchar(10) FK >- Title.tconst
nconst varchar(10) FK >- Actors.nconst

Thank You/Any Questions