Solution – Exercise II

HiveQL, Create and work with External Tables on IMDb Data

Solution

Prerequisites:

- Setup Google Cloud SDK
- Start VM instance
- Pull docker container `marcelmittelstaedt/hive_base:latest`
- Start docker container: `docker run -dit --name hive_base_container -p 8088:8088 -p 9870:9870 -p 9864:9864 marcelmittelstaedt/hive_base:latest`
- Get into docker container
- Start Hadoop and Hive Shell:
  - `start-all.sh`
  - `hive`

Solution

Exercise 1-4:

1. Download and unzip https://datasets.imdbws.com/name.basics.tsv.gz

```
wget https://datasets.imdbws.com/name.basics.tsv.gz
gunzip name.basics.tsv.gz
```

2. Create HDFS directory **/user/hadoop/imdb/name_basics/** for file name.basics.tsv

```
hadoop fs -mkdir /user/hadoop/imdb/name_basics
```

3. Put TSV file to HDFS:

```
hadoop fs -put name.basics.tsv /user/hadoop/imdb/name_basics/name.basics.tsv
```


Solution

Exercise 1-4:

4. Create Hive Table `name_basics`:

```
hive > CREATE EXTERNAL TABLE IF NOT EXISTS name_basics(
    nconst STRING,
    primary_name STRING,
    birth_year INT,
    death_year STRING,
    primary_profession STRING,
    known_for_titles STRING
    ) COMMENT 'IMDb Actors' ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t' ST
ORED AS TEXTFILE LOCATION '/user/hadoop/imdb/name_basics'
TBLPROPERTIES ('skip.header.line.count'='1');
```

Solution

Exercise 5:

a) How many movies and how many TV series are within the IMDB dataset?

```
hive >  SELECT m.title_type, count(*)
        FROM title_basics m GROUP BY m.title_type;

tvMovie 133177
movie 589792
tvEpisode 6107226
tvSeries 216132
[…]

Time taken: 32.908 seconds, Fetched: 12 row(s)
```

b) Who is the youngest actor/writer/… within the dataset?

```
hive >  SELECT * FROM name_basics n
        WHERE n.birth_year = ( SELECT MAX(birth_year) FROM name_basics);
```

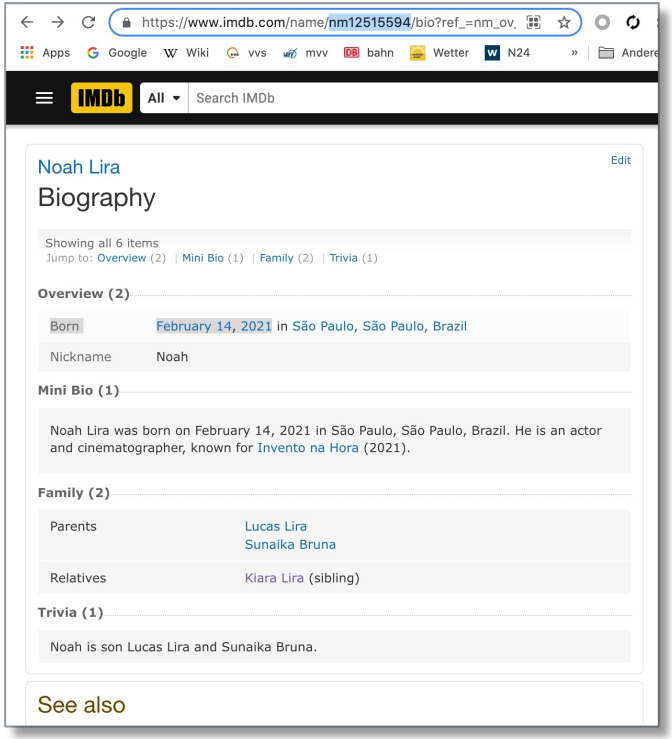
Solution

Exercise 5:

b) Who is the youngest actor/writer/… within the dataset?

```
hive >  SELECT * FROM name_basics n
        WHERE n.birth_year = ( SELECT MAX(birth_year)
                               FROM name_basics);
```

```
nm11495499 Therese Gotlib 2021 NULL producer,camera_department tt14643058
nm12442085 Mae Bair 2021 NULL tt0072506
nm12515594 Noah Lira 2021 NULL actor,cinematographer NULL
nm12533998 Kenna Tota 2021 NULL tt4275008
nm12628453 Rio PenaVega 2021 NULL tt6412982
nm12641996 Lilibet Mountbatten-Windsor 2021 NULL tt0166442
nm12718283 Grace Warrior Irwin Powell 2021 NULL tt14955108,tt8994238,tt0165001
nm12746520 Legend Samuels 2021 NULL tt2224452
nm12919164 Cosmo Jost 2021 NULL NULL
nm9786539 Doguhan Kabadayi 2021 NULL actor tt9873652,tt14858664,tt8309026
Time taken: 65.166 seconds, Fetched: 10 row(s)
```



Solution

Exercise 5:

c) Create a list (`m.tconst, m.original_title, m.start_year, r.average_rating, r.num_votes`) of movies which are:
- equal or newer than year 2010
- have an average rating equal or better than 8,1
- have been voted more than 100.000 times

```
hive >    SELECT m.tconst, m.original_title, m.start_year, r.average_rating, r.num_votes
          FROM title_basics m JOIN title_ratings r on (m.tconst = r.tconst)
          WHERE r.average_rating >= 8.1 and m.start_year >= 2010 and m.title_type = 'movie'
          and r.num_votes > 100000
          ORDER BY r.average_rating desc, r.num_votes DESC;
tt1375666 Inception 2010 8.8 2172818
tt5813916 Dag II 2016 8.7 106823
tt10295212 Shershaah 2021 8.7 102616
tt0816692 Interstellar 2014 8.6 1617714
tt6751668 Gisaengchung 2019 8.6 662581
tt1675434 Intouchables 2011 8.5 797600
tt2582802 Whiplash 2014 8.5 764540
tt1345836 The Dark Knight Rises 2012 8.4 1579445
tt1853728 Django Unchained 2012 8.4 1428387
[…]
```

Solution

Exercise 5:

d) How many movies are in list of c)?

```
hive >    SELECT count(*)
          FROM title_basics m JOIN title_ratings r on (m.tconst = r.tconst)
          WHERE r.average_rating >= 8.1 and m.start_year >= 2010 and m.title_type = 'movie'
          and r.num_votes > 100000;

48
```


Solution

Exercise 5:

e) We want to know which years have been great for cinema.
Create a list with one row per year and a related count of movies which:
- have an average rating better than 8
- have been voted more than 100.000 times
ordered descending by count of movies.

```
hive >  SELECT m.start_year, count(*)
        FROM title_basics m JOIN title_ratings r on (m.tconst = r.tconst)
        WHERE r.average_rating > 8 AND m.title_type = 'movie'
        AND r.num_votes > 100000
        GROUP BY m.start_year
        ORDER BY count(*) DESC;

1995 8
2019 6
2009 6
2016 6
2004 6
2001 6
[…]
```


Solution

Exercise 5:

So 1995 seems to be a really good year for cinema, 8 really good movies have been releases, but which are they?

```
hive > SELECT
            m.tconst, m.original_title, m.start_year, r.average_rating,
            r.num_votes
       FROM title_basics m JOIN title_ratings r ON (m.tconst = r.tconst)
       WHERE
            r.average_rating > 8 AND m.title_type = 'movie'
            AND r.num_votes > 100000 AND m.start_year = 1995
       ORDER BY r.average_rating DESC;

tt0114369 Se7en 1995 8.6 1518316
tt0114814 The Usual Suspects 1995 8.5 1029594
tt0114709 Toy Story 1995 8.3 927722
tt0112573 Braveheart 1995 8.3 989203
tt0113277 Heat 1995 8.2 606627
tt0112641 Casino 1995 8.2 488427
tt0113247 La haine 1995 8.1 160315
tt0112471 Before Sunrise 1995 8.1 287159
[…]
```