**Kindly use the below mentioned Dataset to solve mentioned question and produce your results in pig & hive both.**

**Data Set:** [**Link**](https://drive.google.com/file/d/0ByDwZNhTzoUsVlRmWkNBdzhDV28/view?usp=sharing)

**Dataset Description:**

Column1: Movie ID

Column2: Movie name

Column3: Year of release

Column4: Rating of the movie

Column5: Movie duration in seconds

**Problem Statement:**

A. Find the number of movies released between 1950 and 1960.

B. Find the number of movies having rating more than 4.

C. Find the movies whose rating are between 3 and 4.

D. Find the number of movies with duration more than 2 hours (7200 second).

E. Find the list of years and number of movies released each year.

F. Find the total number of movies in the dataset.

**Solution Required (Pig & Hive Both)**

Hive:

create table movies(movie\_ID int, movie\_name string, year int, rating decimal(10,1), duration int) row format delimited fields terminated by ',';

describe movies;

load data inpath '/movies/hive/moviesdataset.txt' into table movies;

A.Select count(movie\_ID) from movies where year>1950 and year<1960;

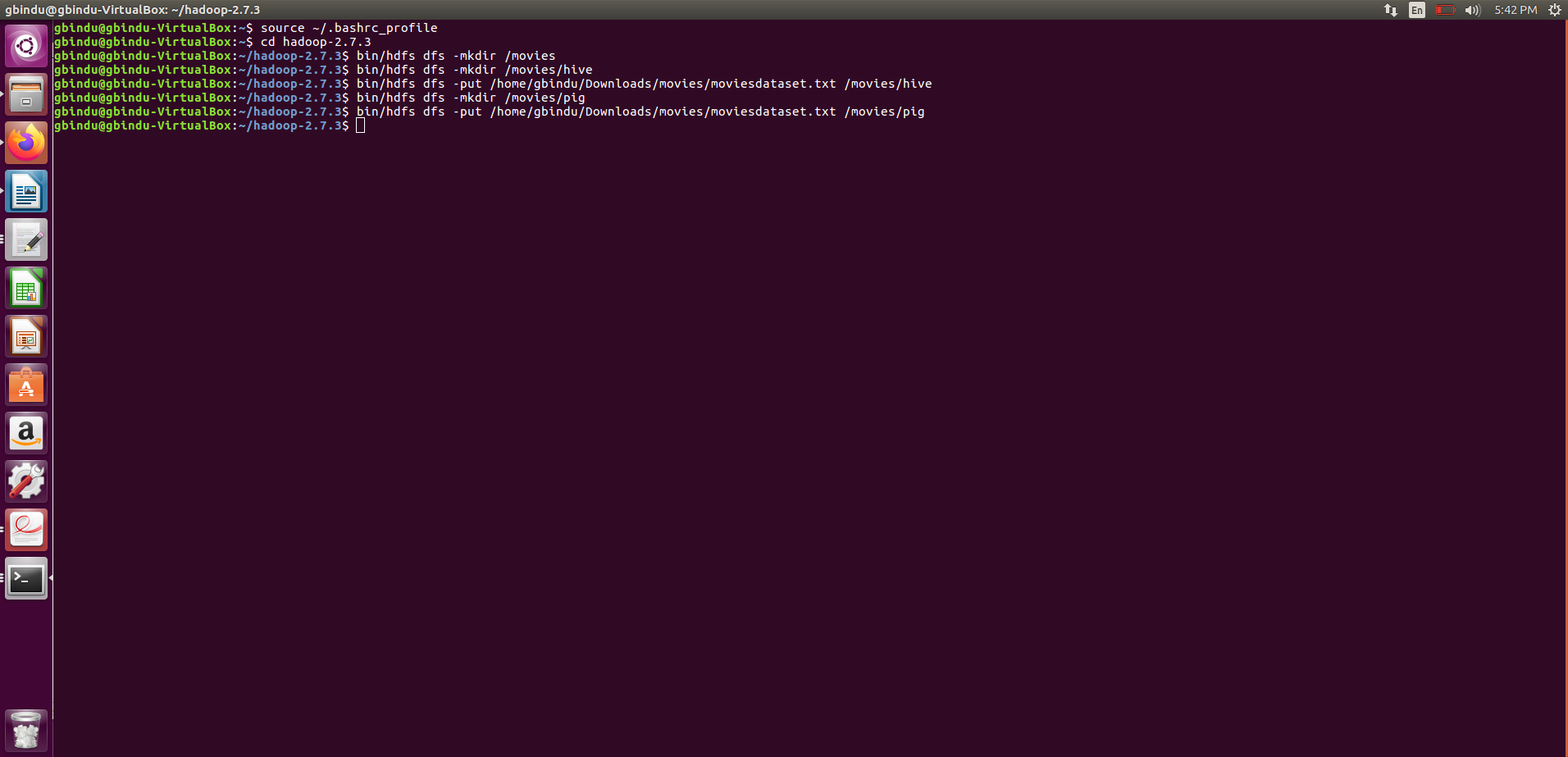
B.Select count(movie\_name) from movies where rating>4.0;

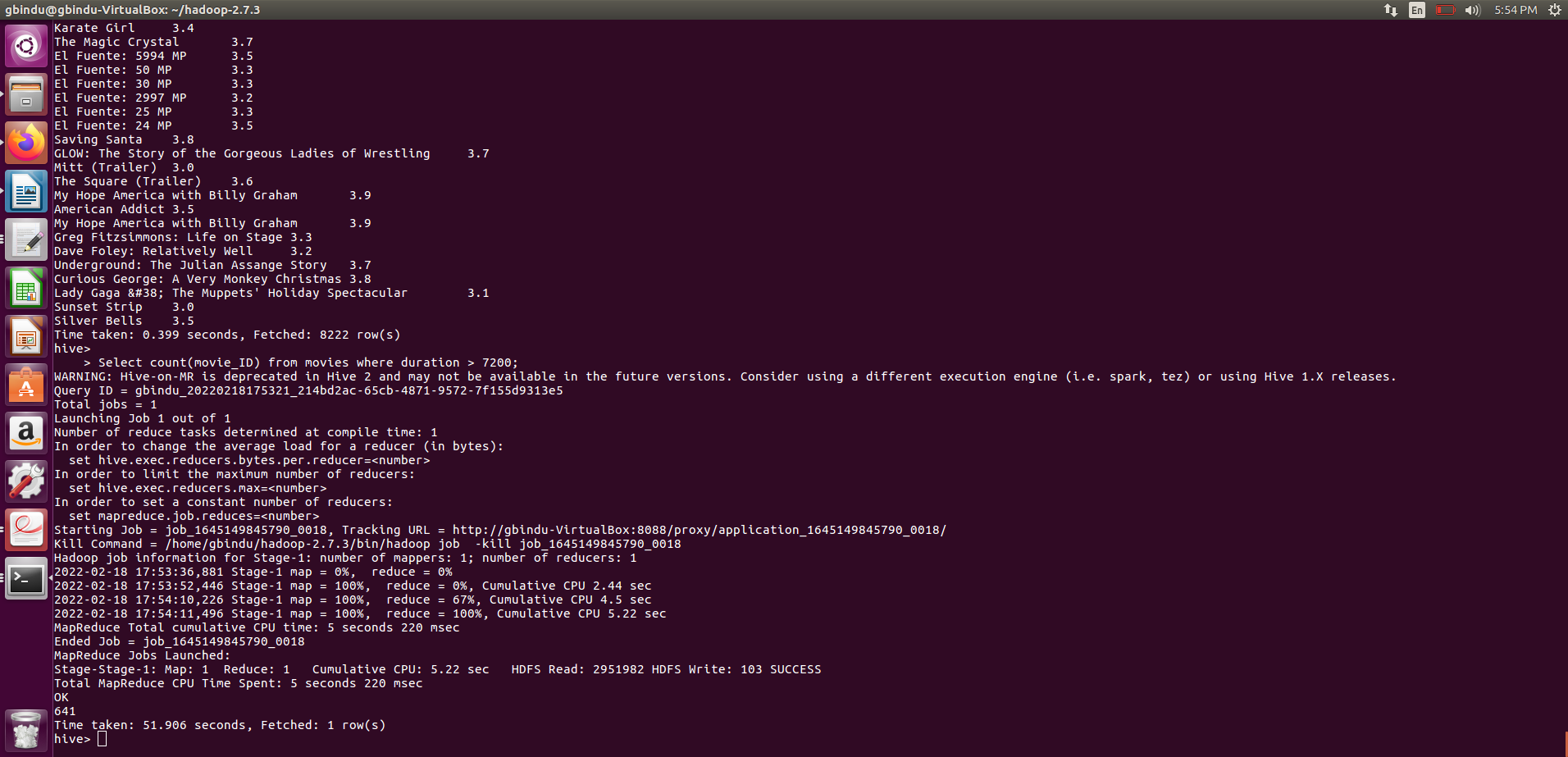
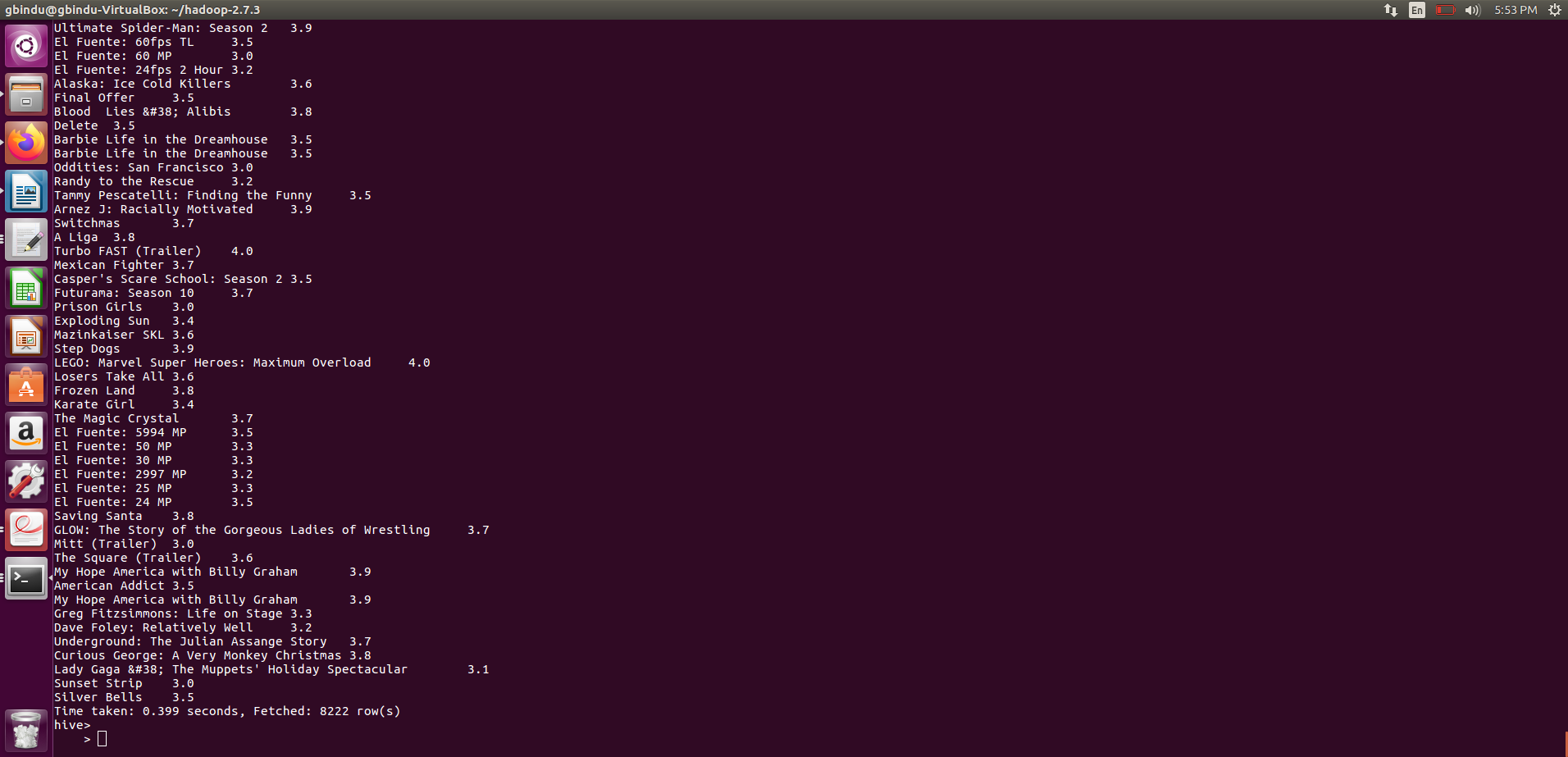
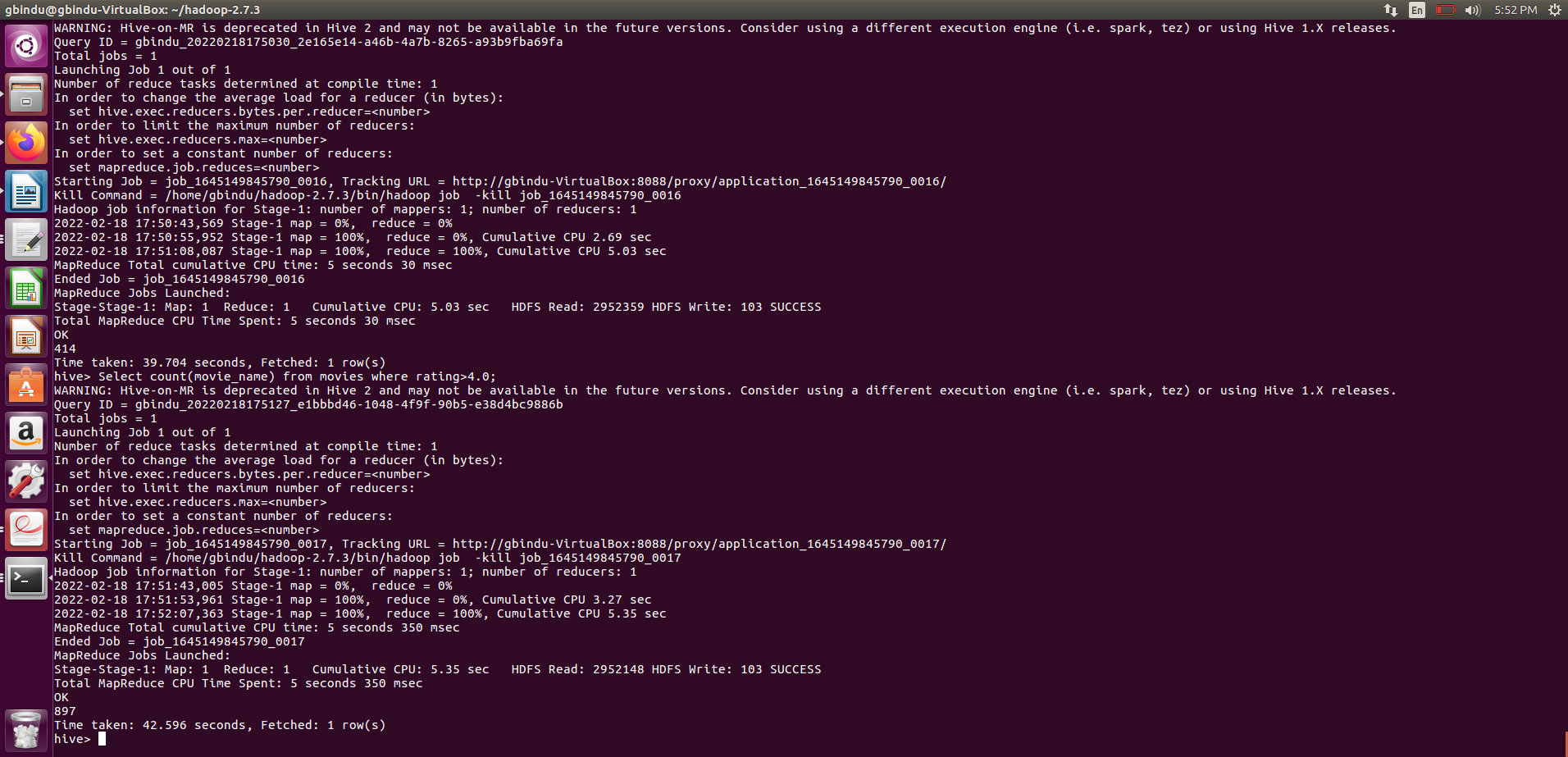
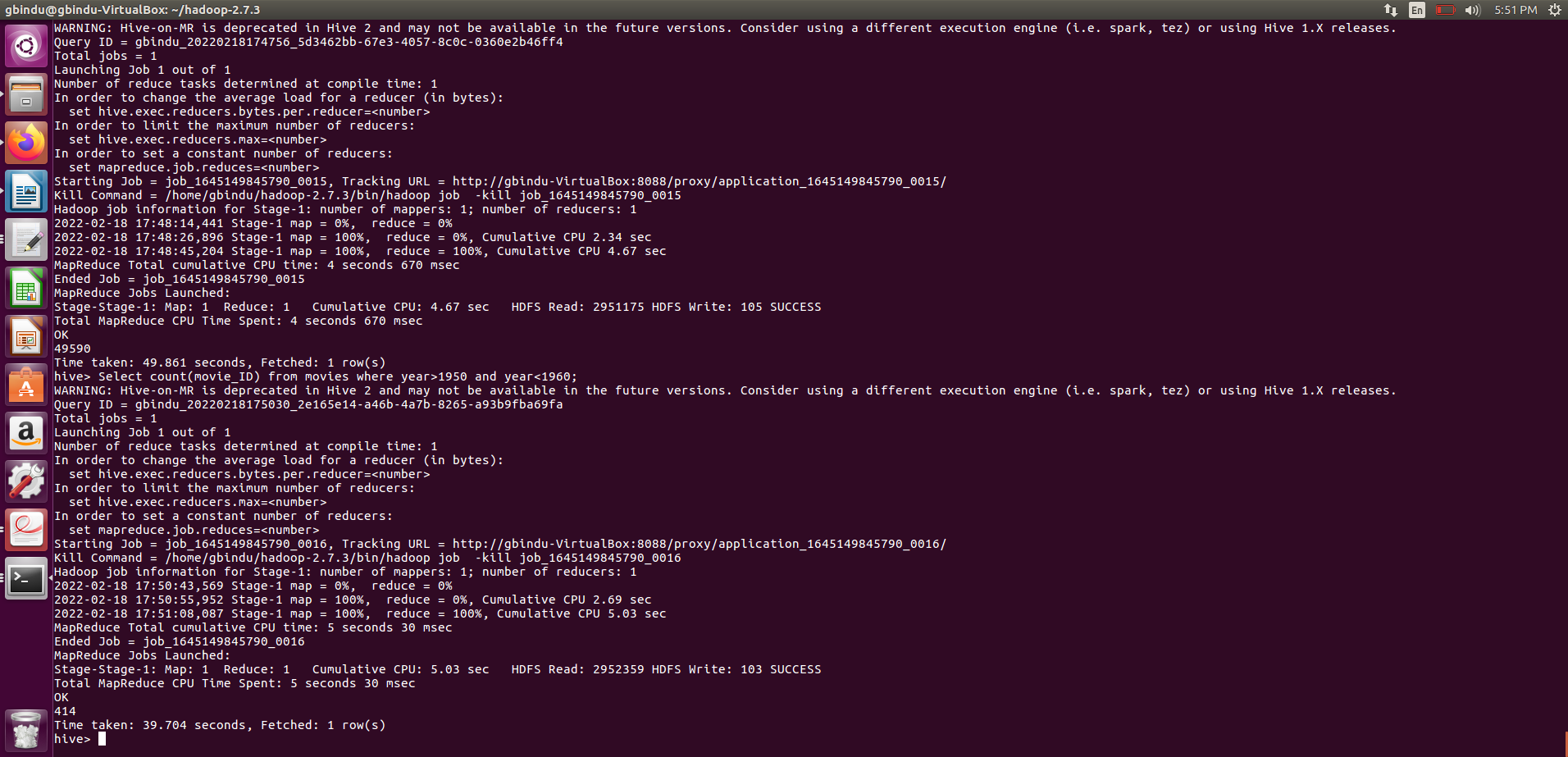
C.Select movie\_name, rating from movies where rating between 3 and 4;

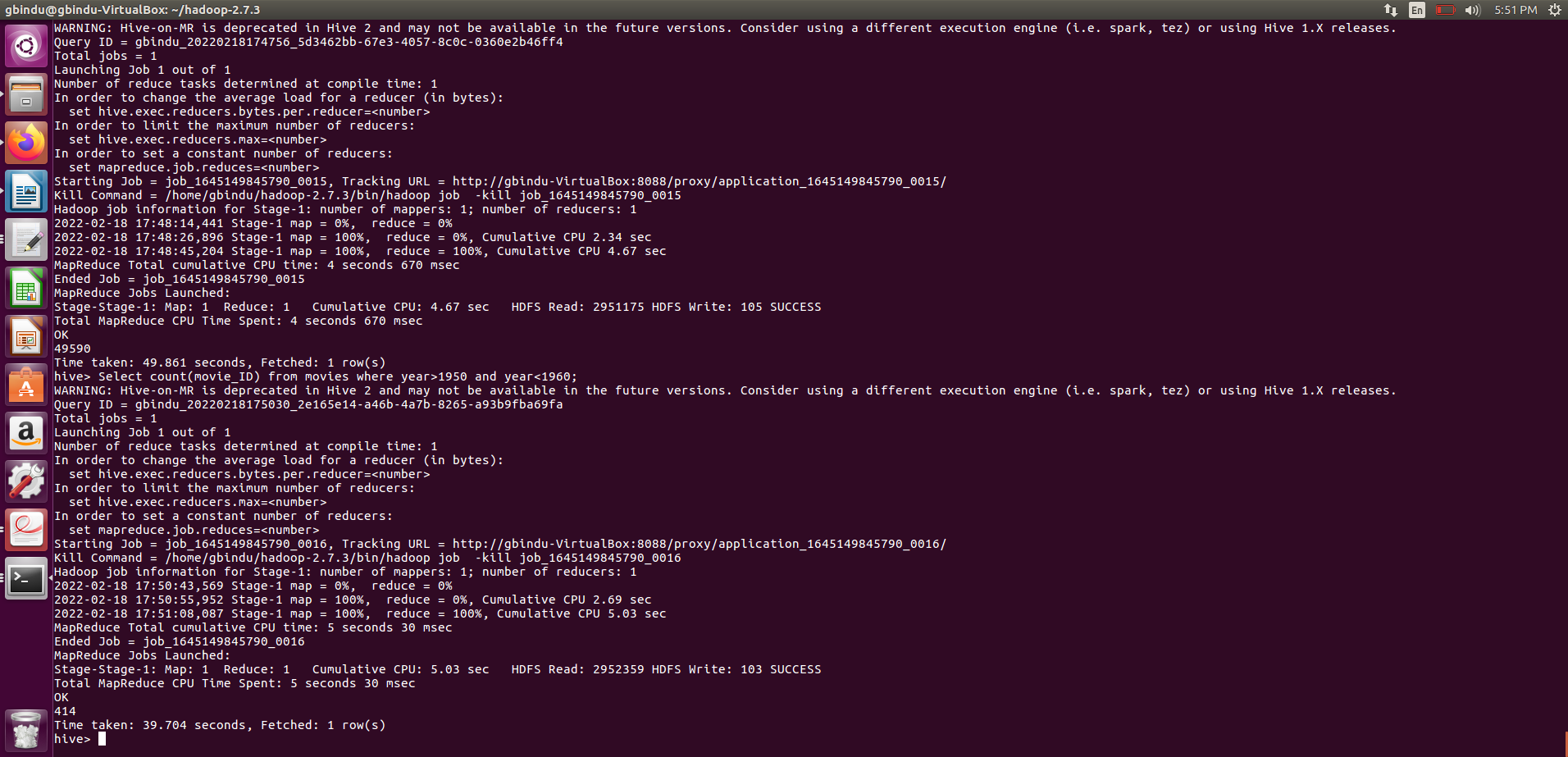
D.Select count(movie\_ID) from movies where duration > 7200;

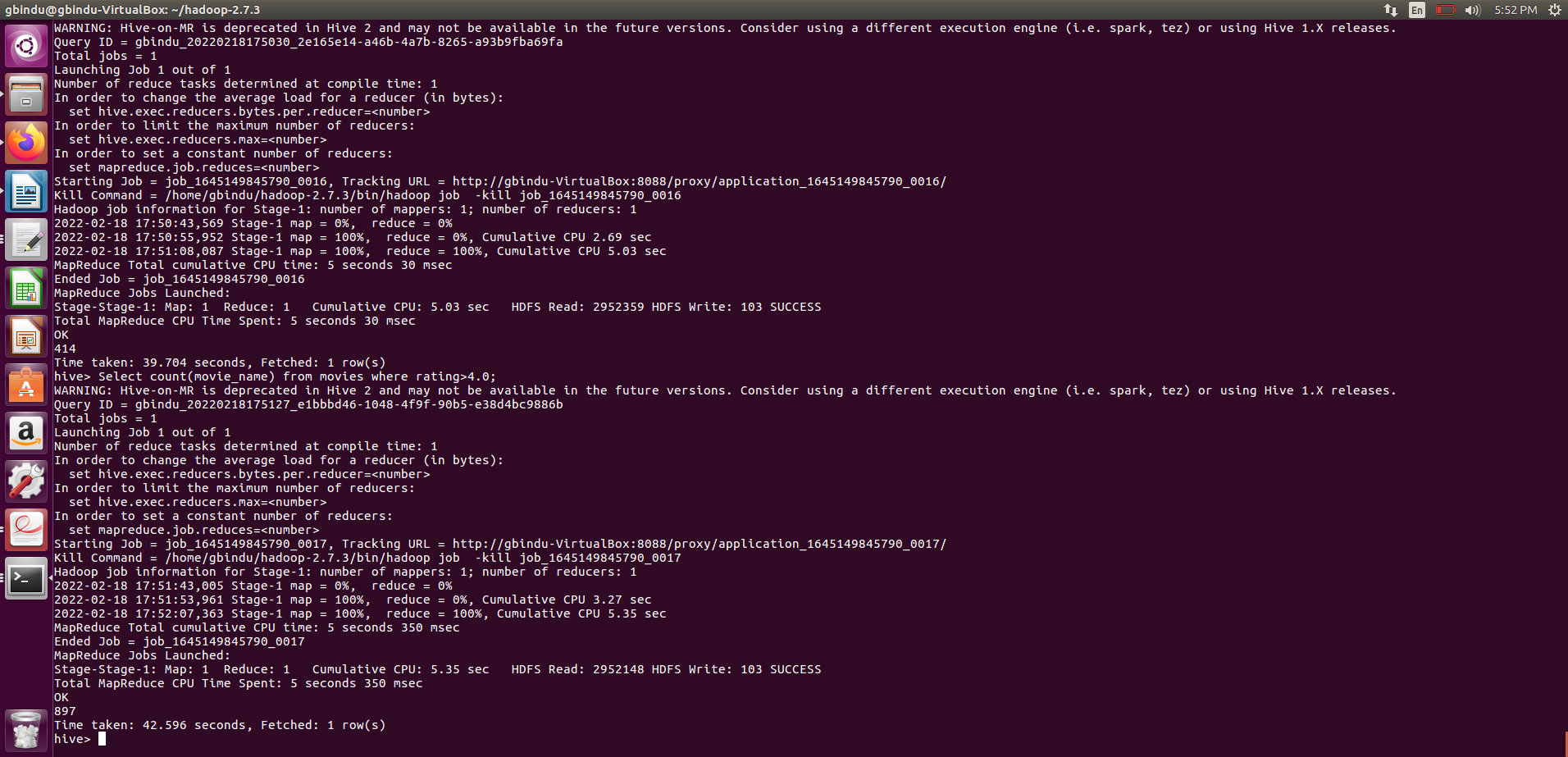
E.Select year, count(movie\_name) from movies group by year;

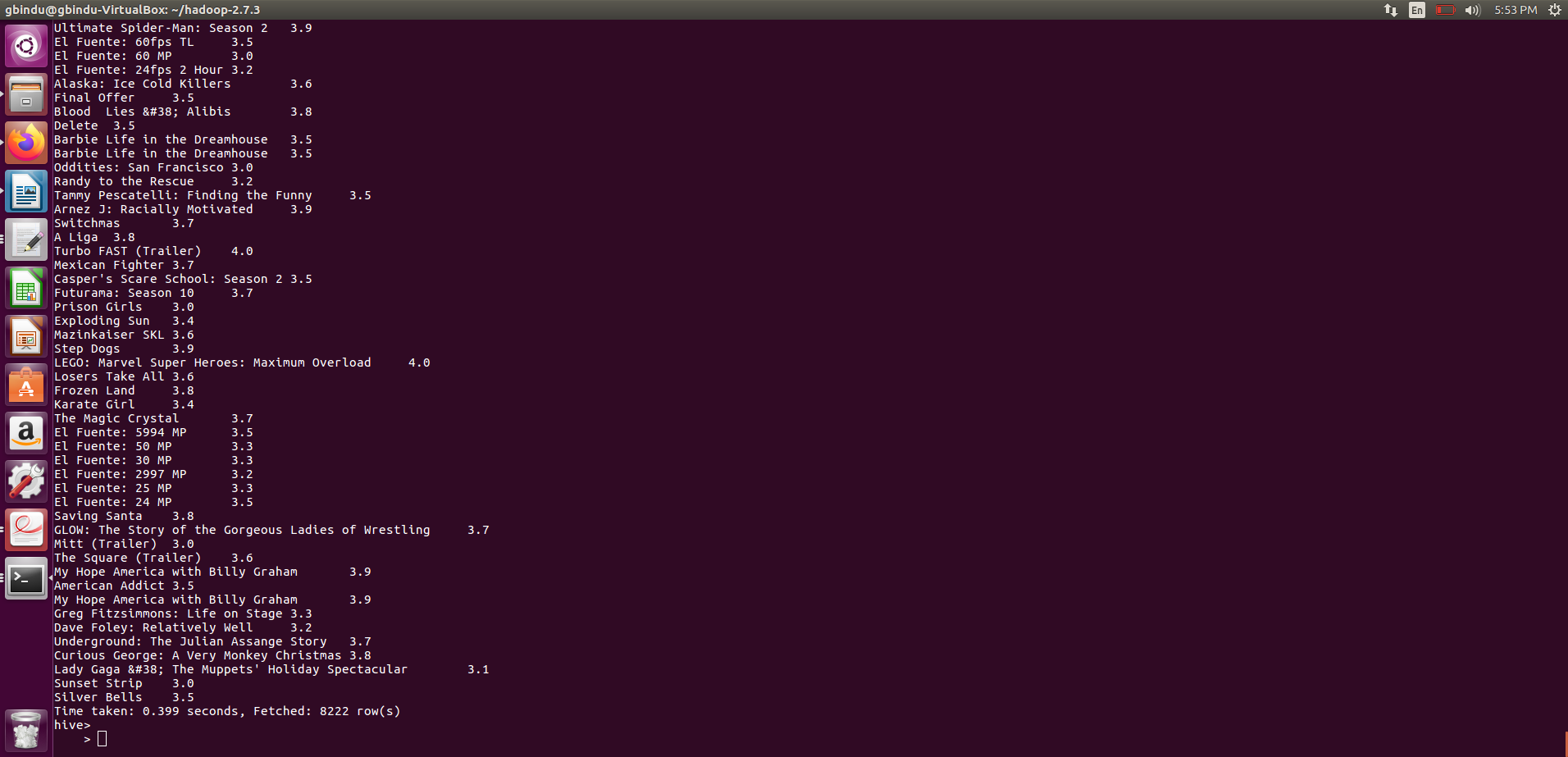
F.Select count(movie\_name) from movies;

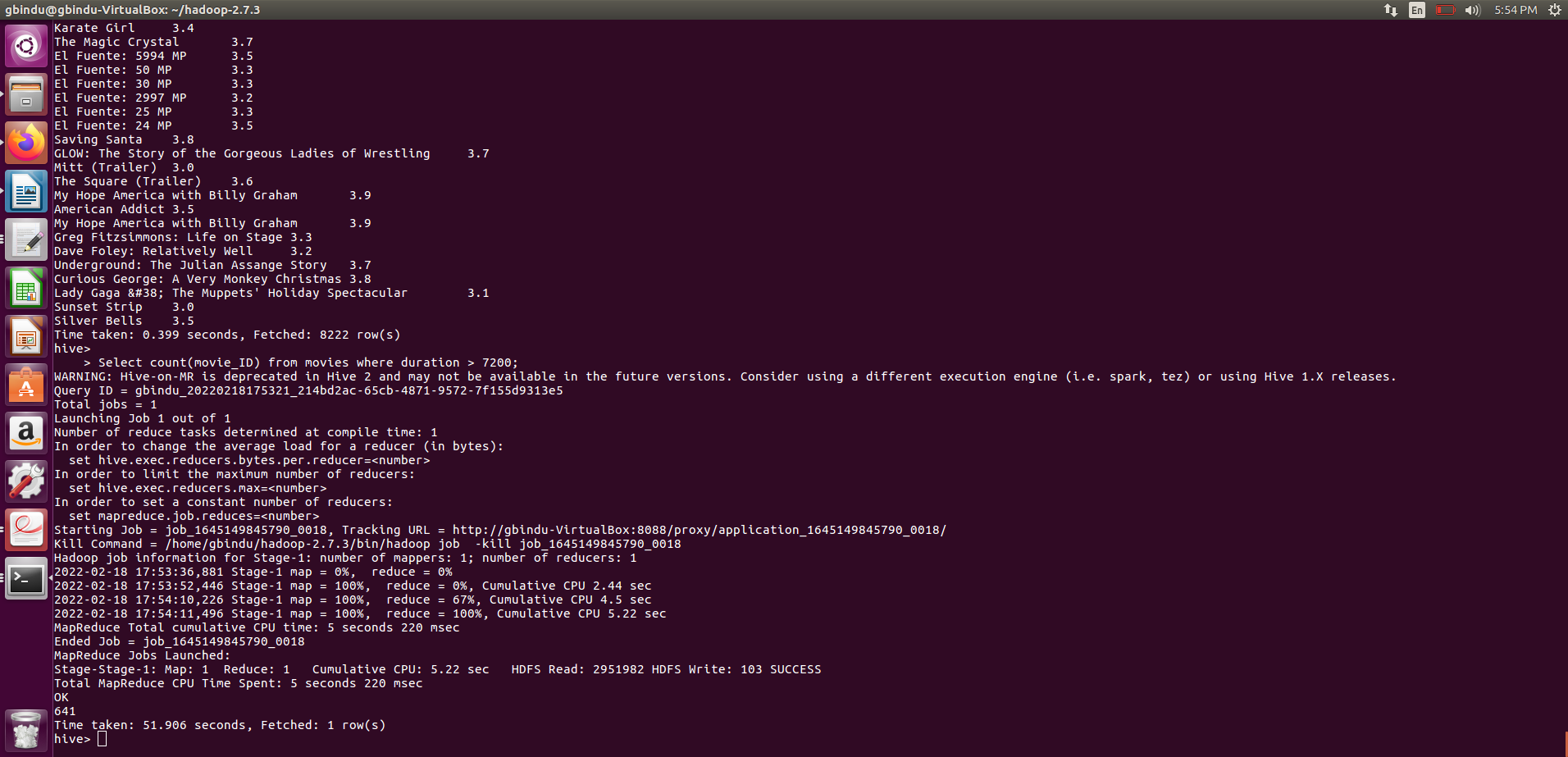
Select count(movie\_name) from movies;



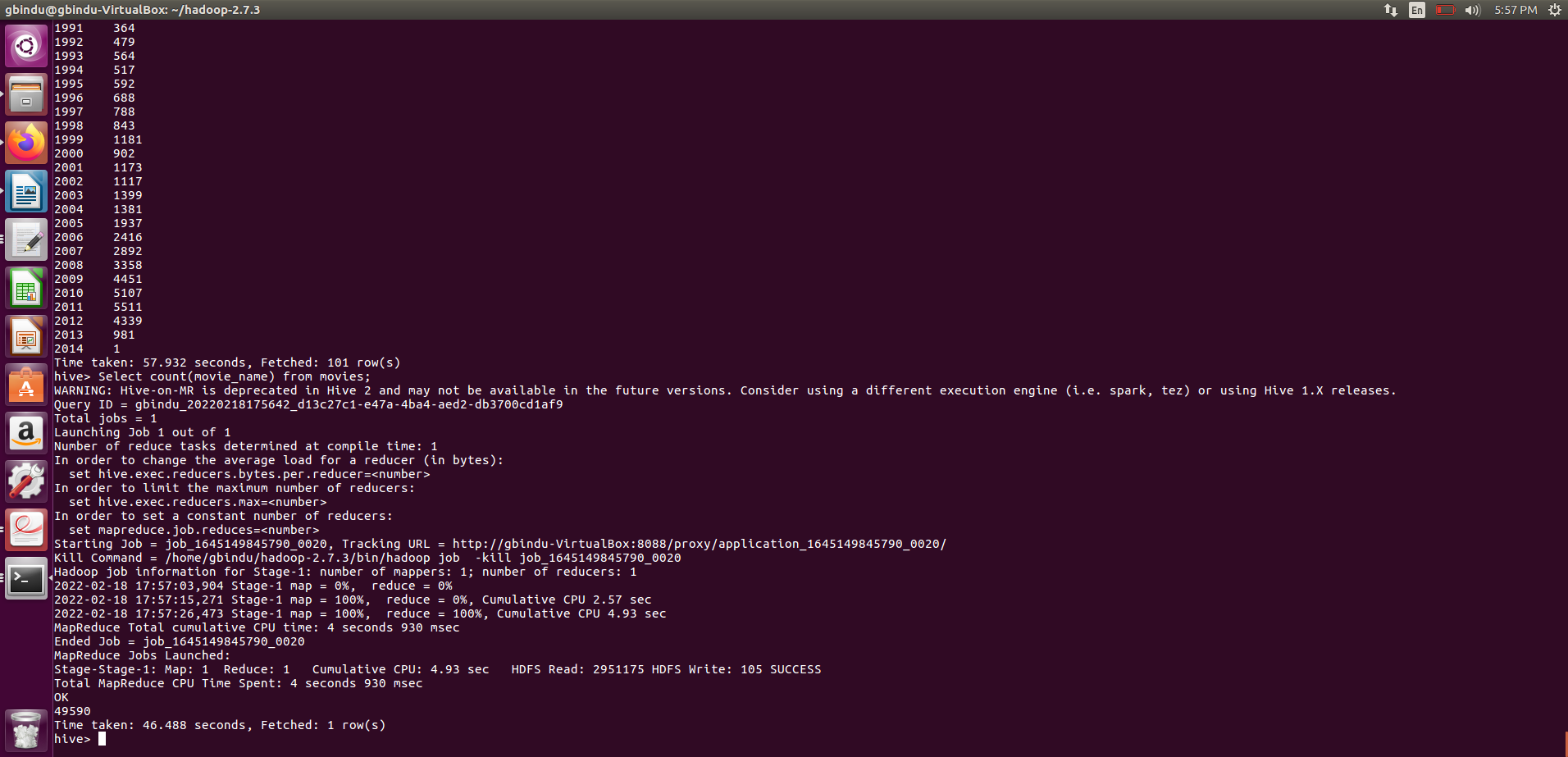












Pig:

A1 = FILTER movies by year>=1950 and year<=1960;

A2 = GROUP A1 ALL;

A3 = FOREACH A2 GENERATE COUNT(A1.movie\_ID);

DUMP A3;

B1 = FILTER movies BY rating >=4.0;

B2 = GROUP B1 ALL;

B3 = FOREACH B2 GENERATE COUNT(B1.movie\_name);

DUMP B3;

C1 = FILTER movies by rating == 4;

C2 = GROUP C1 ALL;

C3 = FOREACH C2 GENERATE GROUP AS movie\_name,rating;

DUMP C3;

D1 = FILTER movies by duration > 7200;

D2 = GROUP D1 ALL;

D3 = FOREACH D2 GENERATE COUNT(D1.movie\_name);

DUMP D3;

E1 =

F1 = FOREACH movies GENERATE COUNT(movie\_name);

DUMP F1;

