**MOVIE DATASET SOLUTION**

**1.Load The File**

**movies = LOAD '/home/cloudera/Desktop/movies.txt' USING PigStorage(',') as (id:int,name:chararray,year:int,rating:int,duration:double);**

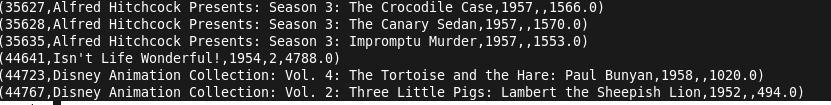
**C:\Users\viks\Desktop\p1.png**

**1. Find the number of movies released between 1950 and 1960.**

**C:\Users\viks\Desktop\p1.png**

**sol.** **FILTER movies by year>1950 and year<1960;**

**o/p**

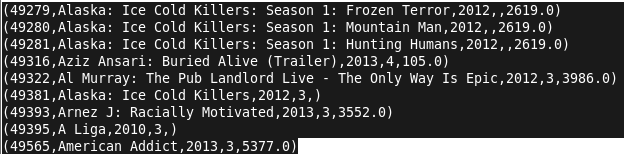
****

**2.** **List the movies that start with the Alpahbet A**

**FILTER movies by name matches 'A.\*';**

**C:\Users\viks\Desktop\p1.png**

**O/P**

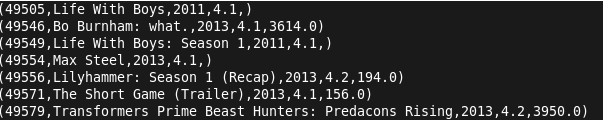
****

**3.** **Find the number of movies having rating more than 4.**

**FILTER movies by rating >4**

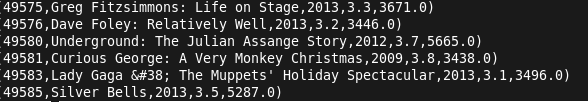
**C:\Users\viks\Desktop\p1.png**

**o/p**

****

**Find the movies whose rating are between 3 and 4**

**a = FILTER movies BY rating>3.0 and rating<4.0;**

****

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

**movies\_duration\_2\_hrs = FILTER movies by duration > 7200;**

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

**grouped\_by\_year = group movies by year;**

**count\_by\_year = FOREACH grouped\_by\_year GENERATE group, COUNT(movies);**

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

**sample\_10\_percent = sample movies 0.1;**

**to check if it is correct or not**

**sample\_group\_all = GROUP sample\_10\_percent ALL;**

**sample\_count = FOREACH sample\_group\_all GENERATE COUNT(sample\_10\_percent.$0);**

**dump sample\_count;**