**Movies Data Analytics**

-- Copy movies dataset to hdfs directory

hdfs dfs -copyFromLocal /home/edureka/Desktop/project/movies.txt /pig

-- Starting Pig

Pig

-- Loading Movies Dataset

MoviesData = LOAD '/pig/movies.txt' using PigStorage(',') AS (custid:int,movie:chararray,year:int,rating:double,duration:int);

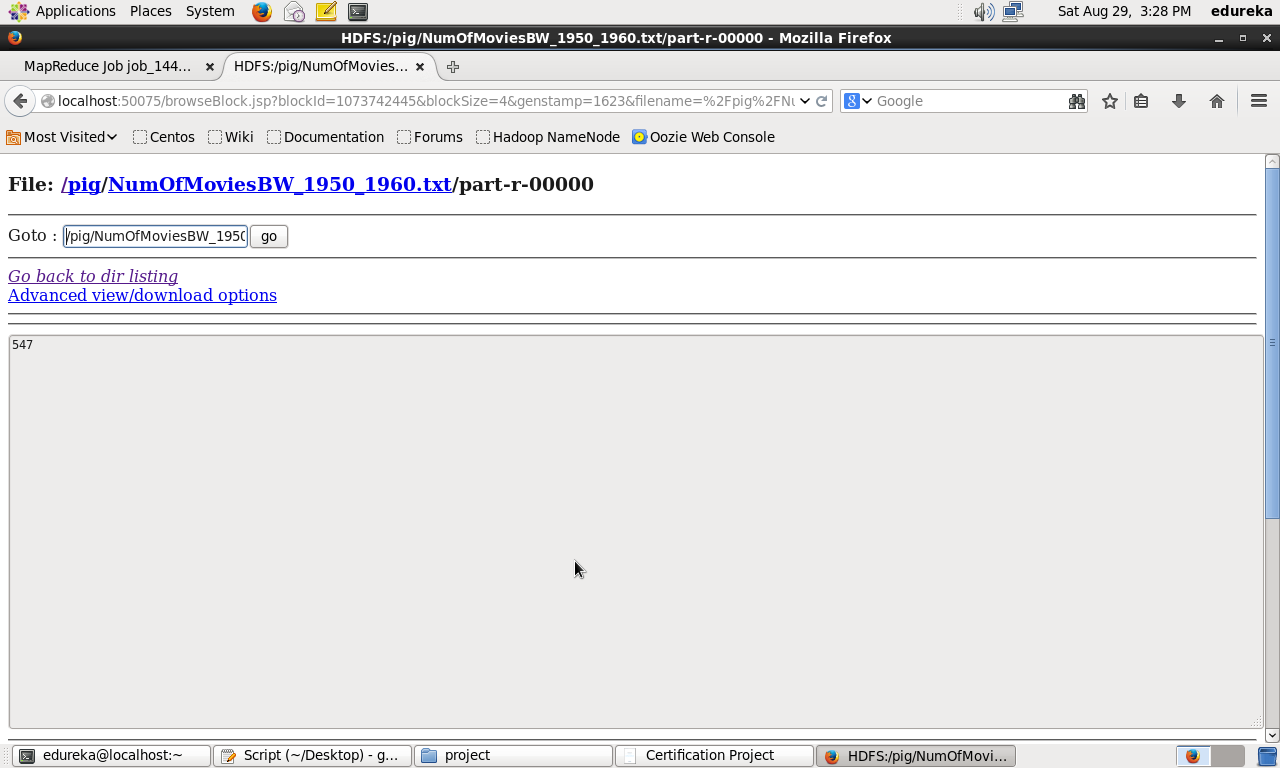
-- To find the number of movies released between 1950 and 1960.

F1 = FILTER MoviesData by ( year >= 1950 ) and ( year <= 1960 );

A = FOREACH ( GROUP F1 ALL ) GENERATE COUNT ( F1 ) ;

--dump A;

store A into '/pig/NumOfMoviesBW\_1950\_1960.txt' using PigStorage();



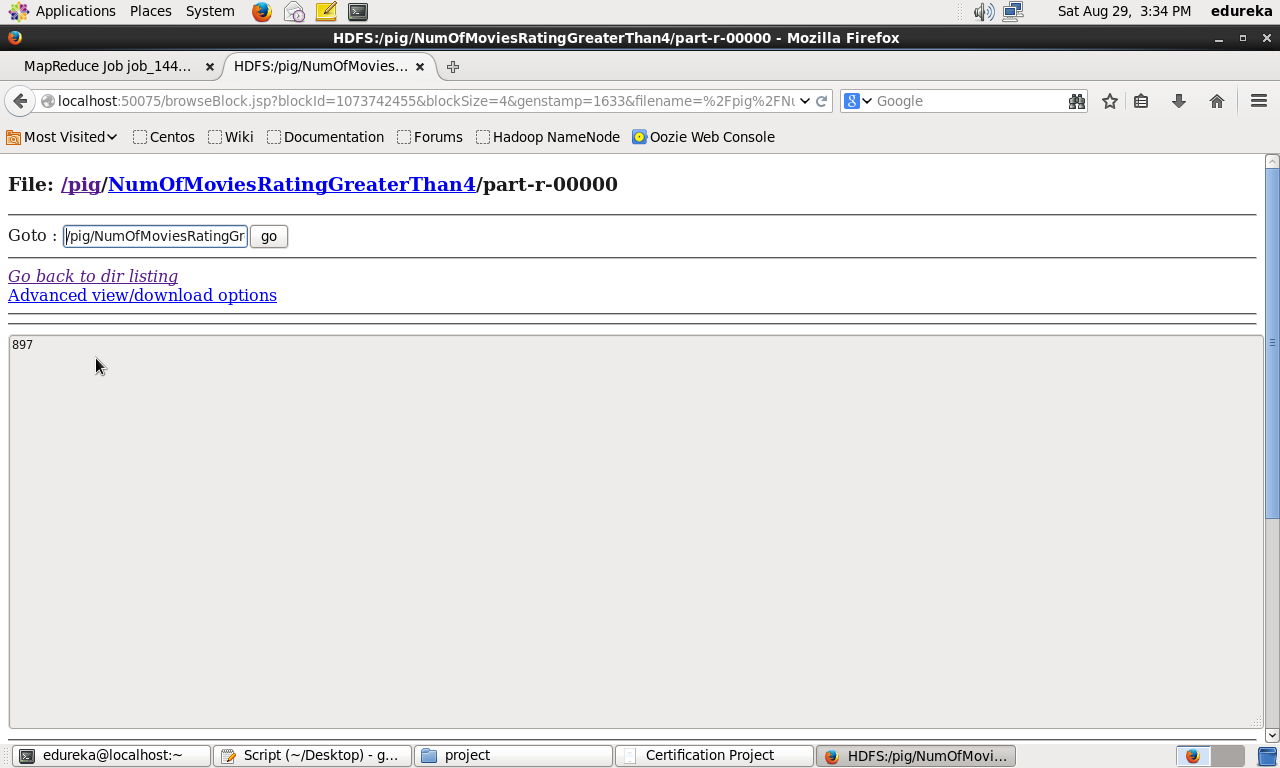
-- Find the number of movies having rating more than 4.

F2 = FILTER MoviesData by ( rating > 4 ) ;

B = FOREACH ( GROUP F2 ALL ) GENERATE COUNT ( F2 ) ;

--dump B;

store B into '/pig/NumOfMoviesRatingGreaterThan4' using PigStorage();



-- Find the movies whose rating is between 3 and 4.

F3 = FILTER MoviesData by ( rating >= 3 AND rating <= 4);

C = FOREACH ( GROUP F3 ALL ) GENERATE COUNT ( F3 ) ;

--dump F3;

store C into '/pig/NumOfMoviesRating\_3\_4' using PigStorage();



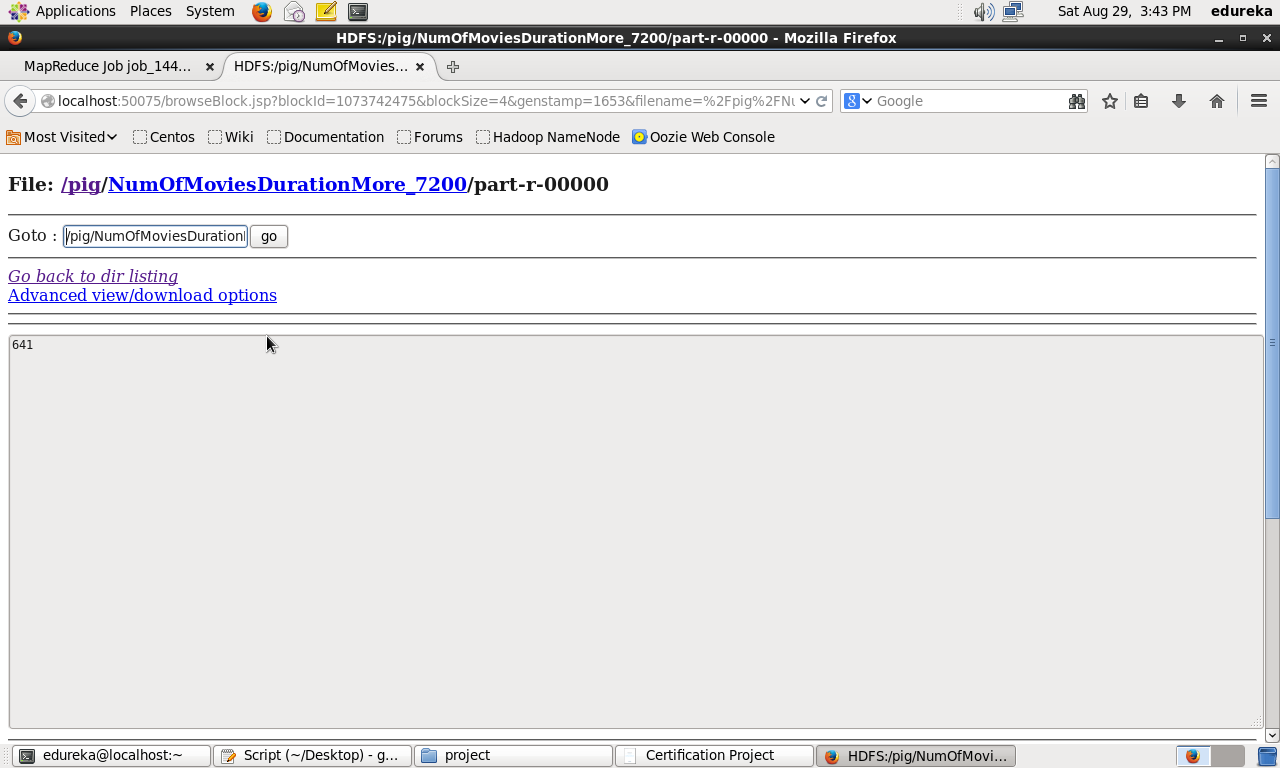
-- Find the number of movies with duration more than 2 hours (7200 second).

F4 = FILTER MoviesData by ( duration > 7200 );

D = FOREACH ( GROUP F4 ALL ) GENERATE COUNT (F4);

--dump D;

store D into '/pig/NumOfMoviesDurationMore\_7200' using PigStorage();



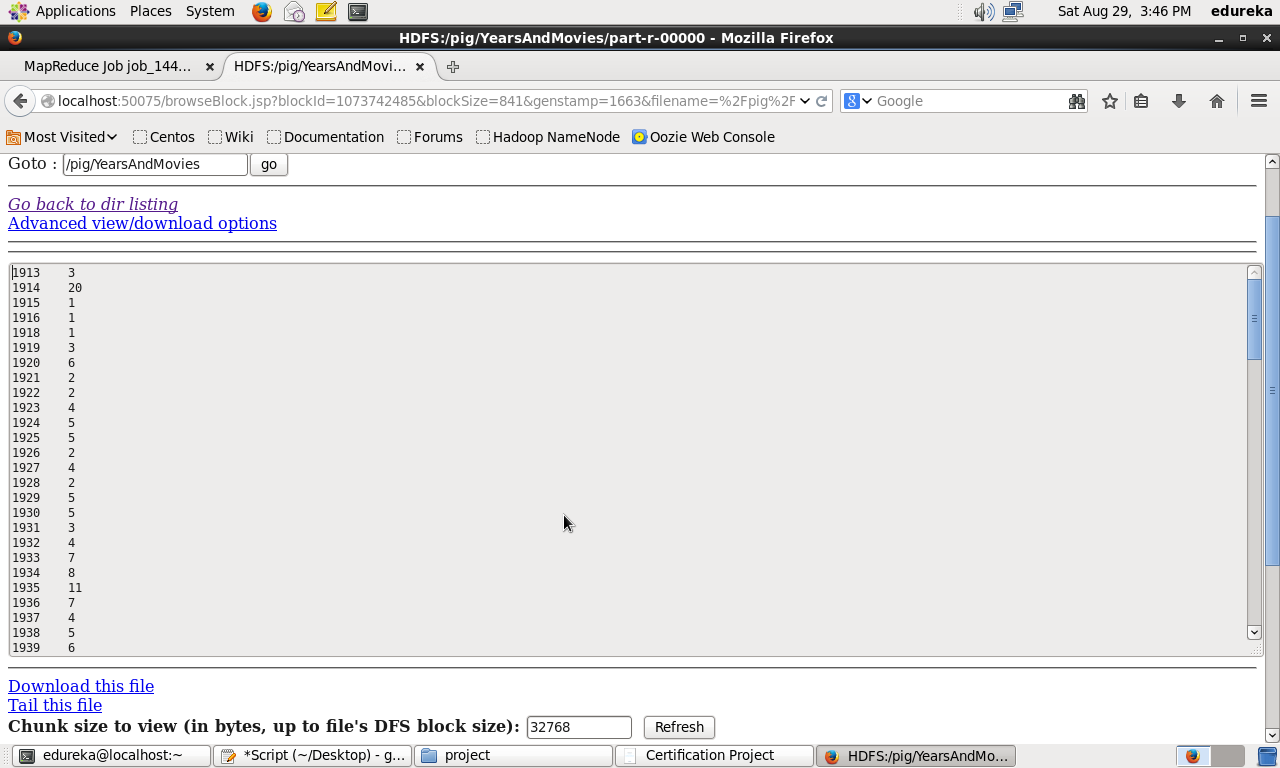
--Find the list of years and number of movies released each year

groupByYear = GROUP MoviesData BY year;

countByYear = FOREACH groupByYear GENERATE group AS Year, COUNT ( MoviesData ) as number;

--dump countByYear;

store countByYear into '/pig/YearsAndMovies' using PigStorage();



Full Solution File

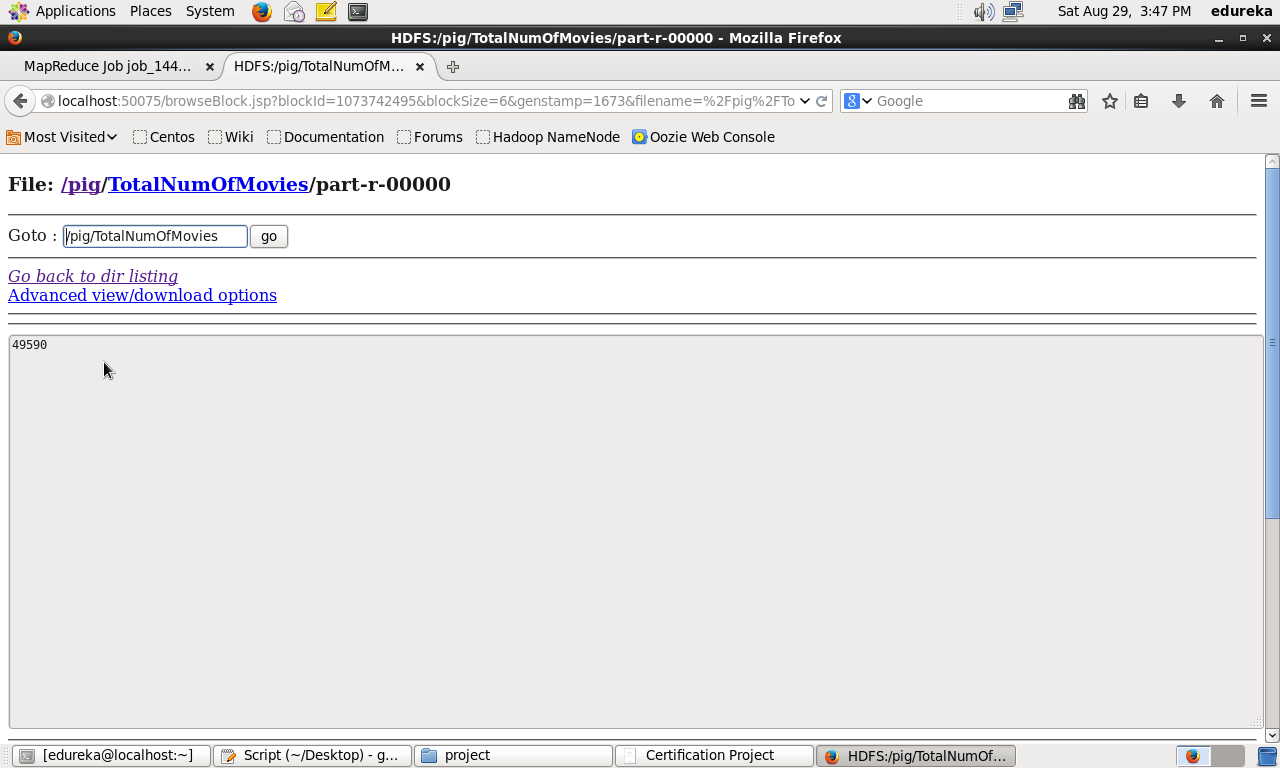


--Find the total number of movies in the dataset.

F = FOREACH ( GROUP MoviesData ALL ) GENERATE COUNT ( MoviesData );

--dump F;

store F into '/pig/TotalNumOfMovies' using PigStorage();



\*\*\*The End\*\*\*