Assignment Week 04

1. Apache Pig group vs co-group

2. Apache Pig Assignment of IMDB data analysis (data will be given).

R1:Load ratings

R2:Load metadata

R3: pick movieID,title & releaseTime (date need to convert into dd-mmm-yyyy) format from R2

R4: group ratings by movieID from R1

R5: take average rating from R4

R6: pick only those from R5 with average rating greater than 4

R7: Join R6 with R3

R8: order by R7 by release time

last step store into HDFS in comma separated file

3. Apache Pig certification https://cognitiveclass.ai/courses/introduction-to-pig

Question 2 Solution but first try yourself:

```
ratings = LOAD '/user/maria_dev/ml-100k/u.data' AS (userID:int, movieID:int, rating:int, ratingTime:int);

metadata = LOAD '/user/maria_dev/ml-100k/u.item' USING PigStorage('|') AS (movieID:int,
movieTitle:chararray, releaseDate:chararray, videoRealese:chararray, imdblink:chararray);

  nameLookup = FOREACH metadata GENERATE movieID, movieTitle,

            ToUnixTime(ToDate(releaseDate, 'dd-MMM-yyyy')) AS releaseTime;

  ratingsByMovie = GROUP ratings BY movieID;

avgRatings = FOREACH ratingsByMovie GENERATE group as movieID, AVG(ratings.rating) as avgRating;

fiveStarMovies = FILTER avgRatings BY avgRating > 4.0;

fiveStarsWithData = JOIN fiveStarMovies BY movieID, nameLookup BY movieID;

oldestFiveStarMovies = ORDER fiveStarsWithData BY nameLookup::releaseTime;

DUMP oldestFiveStarMovies;
```