**Mert Ozer**

**16.10.2017**

**Pig assignment readme file**

**Requirements:**

1. Pig 0.17.0.
2. “Lab2\_extended.pig” file that I provide.
3. Movielens data set. (can be found here: <https://grouplens.org/datasets/movielens/20m/>)
4. Removing headers of used csv files.
5. Changing load lines in “Lab2\_extended.pig” to your path.

**How to run the file?**

bin/pig -x local Lab2\_extended.pig

**Explanation of Map Reduce plans of:**

[**1.1 Show all users which have more than 100 reviews**](#_eln11c3rnect) **1**

[**1.2 Show the total number of reviews for each movie**](#_kgvyq1ecv85h) **3**

[**2.1. The average rating for ’Documentary’ movies**](#_z01rp2krdz46) **4**

[**2.2 For each ’Action’ movie, the total number of tags that have been added**](#_1cffzoa51rjy) **7**

# 1.1 Show all users which have more than 100 reviews

**Phases**

**Map:**  Load A, For each element in C project first element and initial count, Group B

**Combine:**  For each element in C project first element and count, Group B

**Reduce:** Filter and store result

**#--------------------------------------------------**

**# Map Reduce Plan**

**#--------------------------------------------------**

**MapReduce node scope-31**

**Map Plan**

**B: Local Rearrange[tuple]{int}(false) - scope-43**

**| |**

**| Project[int][0] - scope-45**

**|**

**|---C: New For Each(false,false)[bag] - scope-32**

**| |**

**| Project[int][0] - scope-33**

**| |**

**| POUserFunc(org.apache.pig.builtin.COUNT$Initial)[tuple] - scope-34**

**| |**

**| |---Project[bag][1] - scope-35**

**|**

**|---Pre Combiner Local Rearrange[tuple]{Unknown} - scope-46**

**|**

**|---A: New For Each(false,false,false,false)[bag] - scope-15**

**| |**

**| Cast[int] - scope-4**

**| |**

**| |---Project[bytearray][0] - scope-3**

**| |**

**| Cast[int] - scope-7**

**| |**

**| |---Project[bytearray][1] - scope-6**

**| |**

**| Cast[float] - scope-10**

**| |**

**| |---Project[bytearray][2] - scope-9**

**| |**

**| Cast[long] - scope-13**

**| |**

**| |---Project[bytearray][3] - scope-12**

**|**

**|---A: Load(/home/mert/ml-20m/ratings.csv:PigStorage(',')) - scope-2--------**

**Combine Plan**

**B: Local Rearrange[tuple]{int}(false) - scope-47**

**| |**

**| Project[int][0] - scope-49**

**|**

**|---C: New For Each(false,false)[bag] - scope-36**

**| |**

**| Project[int][0] - scope-37**

**| |**

**| POUserFunc(org.apache.pig.builtin.COUNT$Intermediate)[tuple] - scope-38**

**| |**

**| |---Project[bag][1] - scope-39**

**|**

**|---B: Package(CombinerPackager)[tuple]{int} - scope-42--------**

**Reduce Plan**

**D: Store(fakefile:org.apache.pig.builtin.PigStorage) - scope-30**

**|**

**|---D: Filter[bag] - scope-26**

**| |**

**| Greater Than[boolean] - scope-29**

**| |**

**| |---Project[long][1] - scope-27**

**| |**

**| |---Constant(100) - scope-28**

**|**

**|---C: New For Each(false,false)[bag] - scope-25**

**| |**

**| Project[int][0] - scope-20**

**| |**

**| POUserFunc(org.apache.pig.builtin.COUNT$Final)[long] - scope-23**

**| |**

**| |---Project[bag][1] - scope-40**

**|**

**|---B: Package(CombinerPackager)[tuple]{int} - scope-17--------**

**Global sort: false**

# 1.2 Show the total number of reviews for each movie

**Phases**

**Map:**  Load A, foreach countmovie initial count, group movieGroup

**Combine:**  For each element in countmovie project first element and count, Group movieGroup

**Reduce:** count final and store result

**#--------------------------------------------------**

**# Map Reduce Plan**

**#--------------------------------------------------**

**MapReduce node scope-25**

**Map Plan**

**movieGroup: Local Rearrange[tuple]{int}(false) - scope-37**

**| |**

**| Project[int][0] - scope-39**

**|**

**|---countMovie: New For Each(false,false)[bag] - scope-26**

**| |**

**| Project[int][0] - scope-27**

**| |**

**| POUserFunc(org.apache.pig.builtin.COUNT$Initial)[tuple] - scope-28**

**| |**

**| |---Project[bag][1] - scope-29**

**|**

**|---Pre Combiner Local Rearrange[tuple]{Unknown} - scope-40**

**|**

**|---A: New For Each(false,false,false,false)[bag] - scope-13**

**| |**

**| Cast[int] - scope-2**

**| |**

**| |---Project[bytearray][0] - scope-1**

**| |**

**| Cast[int] - scope-5**

**| |**

**| |---Project[bytearray][1] - scope-4**

**| |**

**| Cast[float] - scope-8**

**| |**

**| |---Project[bytearray][2] - scope-7**

**| |**

**| Cast[long] - scope-11**

**| |**

**| |---Project[bytearray][3] - scope-10**

**|**

**|---A: Load(/home/mert/ml-20m/ratings.csv:PigStorage(',')) - scope-0--------**

**Combine Plan**

**movieGroup: Local Rearrange[tuple]{int}(false) - scope-41**

**| |**

**| Project[int][0] - scope-43**

**|**

**|---countMovie: New For Each(false,false)[bag] - scope-30**

**| |**

**| Project[int][0] - scope-31**

**| |**

**| POUserFunc(org.apache.pig.builtin.COUNT$Intermediate)[tuple] - scope-32**

**| |**

**| |---Project[bag][1] - scope-33**

**|**

**|---movieGroup: Package(CombinerPackager)[tuple]{int} - scope-36--------**

**Reduce Plan**

**countMovie: Store(fakefile:org.apache.pig.builtin.PigStorage) - scope-24**

**|**

**|---countMovie: New For Each(false,false)[bag] - scope-23**

**| |**

**| Project[int][0] - scope-18**

**| |**

**| POUserFunc(org.apache.pig.builtin.COUNT$Final)[long] - scope-21**

**| |**

**| |---Project[bag][1] - scope-34**

**|**

**|---movieGroup: Package(CombinerPackager)[tuple]{int} - scope-15--------**

**Global sort: false**

**----------------**

# 2.1. The average rating for ’Documentary’ movies

Job1:

**Phases**

**Map:**  Load ratings, Load movies, filter by a constant, union tuples

**Reduce:** store temporary result

Job2:

**Phases**

**Map:**  Load temporary file, foreach bag calculate avg initial and group by

**Combine:**  For each element in bag calculate avg intermediate and group by

**Reduce:** count avg final and store result

**#--------------------------------------------------**

**# Map Reduce Plan**

**#--------------------------------------------------**

**MapReduce node scope-62**

**Map Plan**

**Union[tuple] - scope-63**

**|**

**|---filteredMoviesWithRating: Local Rearrange[tuple]{int}(false) - scope-33**

**| | |**

**| | Project[int][0] - scope-34**

**| |**

**| |---documentaryMoviesId: New For Each(false)[bag] - scope-14**

**| | |**

**| | Project[int][0] - scope-12**

**| |**

**| |---documentaryMovies: Filter[bag] - scope-8**

**| | |**

**| | Matches - scope-11**

**| | |**

**| | |---Project[chararray][1] - scope-9**

**| | |**

**| | |---Constant(.\*Documentary.\*) - scope-10**

**| |**

**| |---movies: New For Each(false,false)[bag] - scope-7**

**| | |**

**| | Cast[int] - scope-2**

**| | |**

**| | |---Project[bytearray][0] - scope-1**

**| | |**

**| | Cast[chararray] - scope-5**

**| | |**

**| | |---Project[bytearray][1] - scope-4**

**| |**

**| |---movies: Load(/home/mert/ml-20m/movies.csv:PigStorage(',')) - scope-0**

**|**

**|---filteredMoviesWithRating: Local Rearrange[tuple]{int}(false) - scope-35**

**| |**

**| Project[int][1] - scope-36**

**|**

**|---A: New For Each(false,false,false,false)[bag] - scope-28**

**| |**

**| Cast[int] - scope-17**

**| |**

**| |---Project[bytearray][0] - scope-16**

**| |**

**| Cast[int] - scope-20**

**| |**

**| |---Project[bytearray][1] - scope-19**

**| |**

**| Cast[float] - scope-23**

**| |**

**| |---Project[bytearray][2] - scope-22**

**| |**

**| Cast[long] - scope-26**

**| |**

**| |---Project[bytearray][3] - scope-25**

**|**

**|---A: Load(/home/mert/ml-20m/ratings.csv:PigStorage(',')) - scope-15--------**

**Reduce Plan**

**Store(file:/tmp/temp-1991941277/tmp1041519265:org.apache.pig.impl.io.InterStorage) - scope-64**

**|**

**|---filteredMoviesWithRating: New For Each(true,true)[tuple] - scope-47**

**| |**

**| POBinCond[bag] - scope-41**

**| |**

**| |---Project[bag][1] - scope-37**

**| |**

**| |---POUserFunc(org.apache.pig.builtin.IsEmpty)[boolean] - scope-39**

**| | |**

**| | |---Project[bag][1] - scope-38**

**| |**

**| |---Constant({()}) - scope-40**

**| |**

**| POBinCond[bag] - scope-46**

**| |**

**| |---Project[bag][2] - scope-42**

**| |**

**| |---POUserFunc(org.apache.pig.builtin.IsEmpty)[boolean] - scope-44**

**| | |**

**| | |---Project[bag][2] - scope-43**

**| |**

**| |---Constant({(,,,)}) - scope-45**

**|**

**|---filteredMoviesWithRating: Package(Packager)[tuple]{int} - scope-32--------**

**Global sort: false**

**----------------**

**MapReduce node scope-66**

**Map Plan**

**groupped: Local Rearrange[tuple]{int}(false) - scope-79**

**| |**

**| Project[int][0] - scope-81**

**|**

**|---avgMovieRating: New For Each(false,false)[bag] - scope-67**

**| |**

**| Project[int][0] - scope-68**

**| |**

**| POUserFunc(org.apache.pig.builtin.FloatAvg$Initial)[tuple] - scope-69**

**| |**

**| |---Project[bag][3] - scope-70**

**| |**

**| |---Project[bag][1] - scope-71**

**|**

**|---Pre Combiner Local Rearrange[tuple]{Unknown} - scope-82**

**|**

**|---Load(file:/tmp/temp-1991941277/tmp1041519265:org.apache.pig.impl.io.InterStorage) - scope-65--------**

**Combine Plan**

**groupped: Local Rearrange[tuple]{int}(false) - scope-83**

**| |**

**| Project[int][0] - scope-85**

**|**

**|---avgMovieRating: New For Each(false,false)[bag] - scope-72**

**| |**

**| Project[int][0] - scope-73**

**| |**

**| POUserFunc(org.apache.pig.builtin.FloatAvg$Intermediate)[tuple] - scope-74**

**| |**

**| |---Project[bag][1] - scope-75**

**|**

**|---groupped: Package(CombinerPackager)[tuple]{int} - scope-78--------**

**Reduce Plan**

**avgMovieRating: Store(fakefile:org.apache.pig.builtin.PigStorage) - scope-59**

**|**

**|---avgMovieRating: New For Each(false,false)[bag] - scope-58**

**| |**

**| Project[int][0] - scope-52**

**| |**

**| POUserFunc(org.apache.pig.builtin.FloatAvg$Final)[double] - scope-56**

**| |**

**| |---Project[bag][1] - scope-76**

**|**

**|---groupped: Package(CombinerPackager)[tuple]{int} - scope-49--------**

**Global sort: false**

# 2.2 For each ’Action’ movie, the total number of tags that have been added

Job1:

**Phases**

**Map:**  Load tags, group By, Load movies, filter by a constant, union tuples

**Reduce:** store temporary result

Job2:

**Phases**

**Map:**  Load temporary file, foreach bag calculate count initial and group by

**Combine:**  For each element in bag calculate count intermediate and group by

**Reduce:** count final and store result

**#--------------------------------------------------**

**# Map Reduce Plan**

**#--------------------------------------------------**

**MapReduce node scope-60**

**Map Plan**

**Union[tuple] - scope-61**

**|**

**|---actionWtags: Local Rearrange[tuple]{int}(false) - scope-36**

**| | |**

**| | Project[int][0] - scope-37**

**| |**

**| |---actionMoviesWithId: New For Each(false)[bag] - scope-14**

**| | |**

**| | Project[int][0] - scope-12**

**| |**

**| |---actionMovies: Filter[bag] - scope-8**

**| | |**

**| | Matches - scope-11**

**| | |**

**| | |---Project[chararray][1] - scope-9**

**| | |**

**| | |---Constant(.\*Action.\*) - scope-10**

**| |**

**| |---movies: New For Each(false,false)[bag] - scope-7**

**| | |**

**| | Cast[int] - scope-2**

**| | |**

**| | |---Project[bytearray][0] - scope-1**

**| | |**

**| | Cast[chararray] - scope-5**

**| | |**

**| | |---Project[bytearray][1] - scope-4**

**| |**

**| |---movies: Load(/home/mert/ml-20m/movies.csv:PigStorage(',')) - scope-0**

**|**

**|---actionWtags: Local Rearrange[tuple]{int}(false) - scope-38**

**| |**

**| Project[int][1] - scope-39**

**|**

**|---tags: New For Each(false,false,false,false,false)[bag] - scope-31**

**| |**

**| Cast[int] - scope-17**

**| |**

**| |---Project[bytearray][0] - scope-16**

**| |**

**| Cast[int] - scope-20**

**| |**

**| |---Project[bytearray][1] - scope-19**

**| |**

**| Cast[chararray] - scope-23**

**| |**

**| |---Project[bytearray][2] - scope-22**

**| |**

**| Cast[chararray] - scope-26**

**| |**

**| |---Project[bytearray][3] - scope-25**

**| |**

**| Cast[long] - scope-29**

**| |**

**| |---Project[bytearray][4] - scope-28**

**|**

**|---tags: Load(/home/mert/ml-20m/tags.csv:PigStorage(',')) - scope-15--------**

**Reduce Plan**

**Store(file:/tmp/temp-6846399/tmp-1226544847:org.apache.pig.impl.io.InterStorage) - scope-62**

**|**

**|---actionWtags: New For Each(true,true)[tuple] - scope-46**

**| |**

**| POBinCond[bag] - scope-44**

**| |**

**| |---Project[bag][1] - scope-40**

**| |**

**| |---POUserFunc(org.apache.pig.builtin.IsEmpty)[boolean] - scope-42**

**| | |**

**| | |---Project[bag][1] - scope-41**

**| |**

**| |---Constant({()}) - scope-43**

**| |**

**| Project[bag][2] - scope-45**

**|**

**|---actionWtags: Package(Packager)[tuple]{int} - scope-35--------**

**Global sort: false**

**----------------**

**MapReduce node scope-64**

**Map Plan**

**groupped: Local Rearrange[tuple]{int}(false) - scope-76**

**| |**

**| Project[int][0] - scope-78**

**|**

**|---actionWcount: New For Each(false,false)[bag] - scope-65**

**| |**

**| Project[int][0] - scope-66**

**| |**

**| POUserFunc(org.apache.pig.builtin.COUNT$Initial)[tuple] - scope-67**

**| |**

**| |---Project[bag][1] - scope-68**

**|**

**|---Pre Combiner Local Rearrange[tuple]{Unknown} - scope-79**

**|**

**|---Load(file:/tmp/temp-6846399/tmp-1226544847:org.apache.pig.impl.io.InterStorage) - scope-63--------**

**Combine Plan**

**groupped: Local Rearrange[tuple]{int}(false) - scope-80**

**| |**

**| Project[int][0] - scope-82**

**|**

**|---actionWcount: New For Each(false,false)[bag] - scope-69**

**| |**

**| Project[int][0] - scope-70**

**| |**

**| POUserFunc(org.apache.pig.builtin.COUNT$Intermediate)[tuple] - scope-71**

**| |**

**| |---Project[bag][1] - scope-72**

**|**

**|---groupped: Package(CombinerPackager)[tuple]{int} - scope-75--------**

**Reduce Plan**

**actionWcount: Store(fakefile:org.apache.pig.builtin.PigStorage) - scope-57**

**|**

**|---actionWcount: New For Each(false,false)[bag] - scope-56**

**| |**

**| Project[int][0] - scope-51**

**| |**

**| POUserFunc(org.apache.pig.builtin.COUNT$Final)[long] - scope-54**

**| |**

**| |---Project[bag][1] - scope-73**

**|**

**|---groupped: Package(CombinerPackager)[tuple]{int} - scope-48--------**

**Global sort: false**

**----------------**