Session 4: Schedulers in YARN & Introduction to Pig

Assignment 4.2

Create a sample dataset and implement the below Pig commands on the same dataset.

1) Concat

2) Tokenize

3) Sum

4) Min

5) Max

6) Limit

7) Store

8) Distinct

9) Flatten

10) IsEmpty

Solution:

The file lists movie names and its details like release year, rating and runtime.

File : 

Loaded the file in Pig using following command in grunt shell

movies = LOAD 'SampleDataSet.csv' USING PigStorage(',') as (mov\_id:int,mov\_name:chararray,year:int,rating:double,duration:long);

Running commands on the same:-

1. **Concat**





1. **Tokenize**





1. **Sum**





1. **Min**





1. **Max**





1. **Limit**





1. **Store**

grunt> sorted\_movies = ORDER movies by year;

grunt> STORE sorted\_movies INTO ' Assignment4\_2' USING PigStorage (',');

Input(s):

Successfully read 10 records from: "file:///home/acadgild/Acadgild/Assignment4\_2/SampleDataSet.csv"

Output(s):

Successfully stored 10 records in: "file:///home/acadgild/Acadgild/Assignment4\_2/Assignment4\_2"

Counters:

Total records written : 10

Total bytes written : 0

Spillable Memory Manager spill count : 0

Total bags proactively spilled: 0

Total records proactively spilled: 0



1. **Distinct**

Copied all the row in data set and loaded it into PIG

Grunt> movies = LOAD 'SampleDataSet\_copy.txt' USING PigStorage(',') as (mov\_id:int,mov\_name:chararray,year:int,rating:double,duration:long);

Grunt> dump movies; -- returns 20 rows

grunt> unique\_movies = DISTINCT movies;

grunt> dump unique\_movies;



1. **Flatten**

grunt> filtered\_movies = FILTER movies by year == 1994;

grunt> movie\_group = group filtered\_movies by year;

grunt> dump movie\_group;



grunt> flatten\_example = foreach movie\_group generate group, flatten(filtered\_movies);

grunt> dump flatten\_example

(1994,8,Mother's Boys,1994,3.4,5733)

(1994,7,Muriel's Wedding,1994,3.5,6323)



1. **IsEmpty**

Created new data set 'SampleDataSet\_WithEmpty.txt’ and loaded it into PIG.

grunt> movies1 = LOAD 'SampleDataSet.csv' USING PigStorage(',') as (mov\_id:int,mov\_name:chararray,year:int,rating:double,duration:long);

grunt> dump movies1;



grunt> movies2 = LOAD 'SampleDataSet\_WithEmpty.txt' USING PigStorage(',') as (mov\_id:int,mov\_name:chararray,year:int,rating:double,duration:long);

grunt> dump movies2;



grunt> movie\_joined = cogroup movies1 by mov\_id , movies2 by mov\_id;

grunt> dump movie\_joined;



grunt> movies\_filtered1 = filter movie\_joined by IsEmpty(movies2);

grunt> dump movies\_filtered1;



grunt> movies\_filtered2 = filter movie\_joined by not IsEmpty(movies2);

grunt> dump movies\_filtered2;

