Youtube's data analysis using Pig

Link for dataset: https://edureka.wistia.com/medias/6cchxi6to4

Dataset description:

Column1: Video id of 11 characters.

Column2: uploader of the video of string data type.

Column3: Interval between day of establishment of Youtube and the date of uploading of the video of integer data type.

Column4: Category of the video of String data type.

Column5: Length of the video of integer data type.

Column6: Number of views for the video of integer data type.

Column7: Rating on the video of float data type.

Column8: Number of ratings given on the video.

Column9: Number of comments on the videos in integer data type.

Column10: Related video ids with the uploaded video.

Problem statement:

1) Find out the top 5 categories with maximum number of videos uploaded.

2) Find out the top 10 rated videos.

3) Find out the most viewed videos.

Script for 1st problem statement:

1. youtube = load 'youtubedata.txt' using PigStorage();

2. youtube_col_idandcategory = foreach youtube generate $0 as id ,$3 as category;

3. group_data = group youtube_col_idandcategory by category;

4. count_category = foreach group_data generate group,
COUNT(youtube_col_idandcategory.category) as category_count;

5. Sort_data = order count_category by category_count desc;

6. final_result = limit Sort_data 5;

7.dump final_result;

Explanation for the above scripts:

1. Input file is loaded from HDFS into table named youtube.

2. 1st and 4th column i.e, video and category are extracted from the youtube table and stored in a new table named youtube_col_idandcategory.

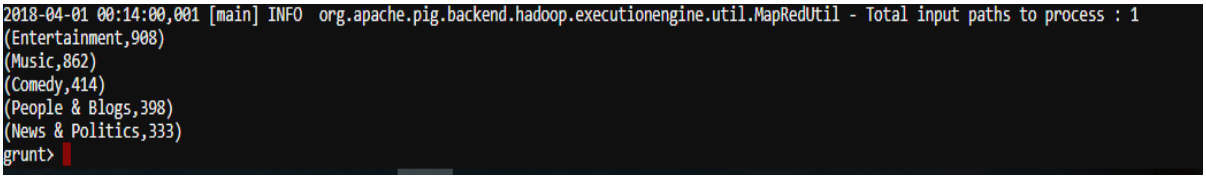
3. The entries from the above table i.e youtube_col_idandcategory are grouped according to category column.

4. From the youtube_col_idandcategory's category, the count of videos from every category are generated and stored in table named count_category.

5. The entries from the table count_category are arranged in a descending order of the number of videos.

6. The top 5 entries are selected and it gives the desired result.

Output:



Scripts for the 2nd problem statement:

1. youtube = load 'youtubedata.txt' using PigStorage();

2. youtube_col_idandrating = foreach youtube generate $0 as id, $6 as rating;

3. order_data = order youtube_col_idandrating by rating desc;

4. final_data = limit order_data 10;

5. dump final_data;

Explanation for the above scripts:

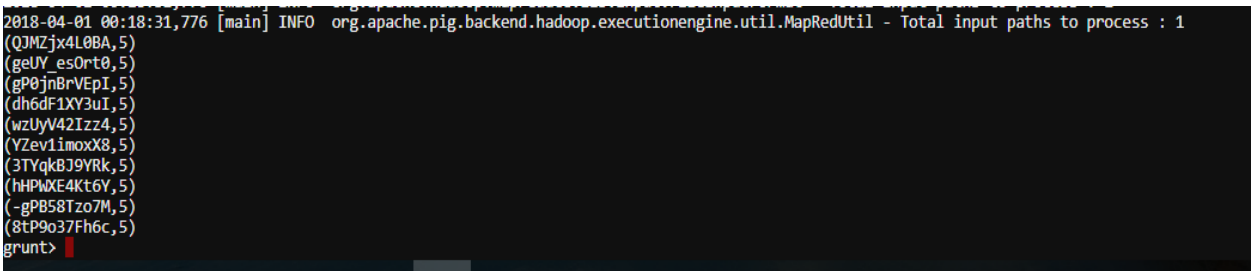
1. Dataset is loaded from HDFS into the table named youtube.

2. 1st and 7th column i.e video and rating of that video are extracted from the youtube table and stored in a new table named youtube_col_idandrating.

3. The youtube_col_idandrating table is arranged in descending order of ratings and stored in the table named order_data.

4. The 10 top entries are extracted and stored in the table named final_data to achieve the final result.

Output:



Scripts for 3rd problem statement:

1. youtube = load 'youtubedata.txt' using PigStorage();

2. youtube_col_idandview= foreach youtube generate $0 as id,(int)$5 as view;

3. youtube_final = filter youtube_col_idandview by $1 is not null;

4. order_data = order youtube_final by view desc;

5. final_data = limit order_data 10;

6. dump final_data;

Explanation for the above scripts:

1. Dataset is loaded from HDFS into the table named youtube.

2. 1st and 6th column i.e video and number of views are extracted from the youtube table and stored into a new table called youtube_col_idandview.

3. From relation youtube_col_idandview the entries with number of views as null are rejected and filtered values are stored in table named youtube_final.

4. The records in the table youtube_final are arranged in descending order of number of views and stored in a new table called order_data.

5. The top 10 entries are extracted and stored in the table final_data and final result is achieved by dumping table final_data.

Output: