Movie Revenue Analytics Data Ingestion
Realtime and Big Data Analytics

Ishita Jaisia [ij2056]

Dataset

The primary source of this dataset is www.themoviedb.org. This dataset contains information on various directors, awards for which they were nominated, the number of total awards won, etc. This dataset consists of 8 files with varying numbers of records. The files we will be using from this dataset are 900_acclaimed_directors_awards.csv and 220k_awards_by_directors.csv.

- *900_acclaimed_directors_awards.csv:* This presents us with details about different directors and the number of awards they received.
  Number of columns : 6676
  Number of rows : 894
- *220k_awards_by_directors.csv:* This consists of details about the award nominations that different directors received in various award ceremonies.
  Number of columns : 6
  Number of rows : 225676

Changed the file from csv format to tsv and processed the data accordingly because the value in a cell might contain a “,”.

First, we perform the below mentioned steps on **900_acclaimed directors_awards.tsv**

Profiling

Rows :
ROW_COUNT = 893

DirectorName :
NOT_NULL = 893
UNIQUE = 893

Total Awards :
NOT_NULL = 893

tmdbID :
NOT_NULL = 893
UNIQUE = 893

Cleaning

Since all the rows are the required format and none of them have any unexpected data we do not need to filter any bad rows from this file.

Extraction
A new table was created which contains a subset of the 6676 columns. In the new table, we have DirectorName, tmdbId and the Total Number of Awards the director won.

Secondly, we perform the below mentioned steps on **220k_awards_by_directors.tsv**

Profiling
Rows :
ROW_COUNT = 225675
DirectorName :
NOT_NULL = 225675
UNIQUE = 29498
DUPLICATE = 196177
Award Ceremony :
NOT_NULL = 225675
Award Year:
NOT_NULL = 225675
Outcome :
NOT_NULL = 225675

Cleaning
Since all the rows are the required format and none of them have any unexpected data we do not need to filter any bad rows from this file.

Extraction
Columns were filtered and a new table was created with the columns DirectorName and a list of Years in which they were nominated for an award in the “Cannes Film Festival”.

Once this was done, a mapping from DirectorName to the years they have been nominated for the award has been created.

Both the files created will be used for the hypothesis where we determine whether the popularity of a director in a particular calendar year affects the chances of them being nominated for an award at the Cannes Film Festival.

Steps to run the code
1. Login to peel cluster
2. Navigate to the respective directory
3. Run the following command hadoop jar <jarFile> <Class with main method> <data input path on HDFS> <output directory>