Movie Revenue Analytics
Realtime and Big Data Analytics

FNU shivanshi [ss14396]

Overview

As the movie industry grows, the probable profit made by a movie becomes of utmost importance for the stakeholders. Among the movies produced between 2010 to 2020 in the United States, less than 40% of the movies had revenues higher than the production budget. This highlights the importance of knowing the factors contributing to the profitability of a movie to make the right investment decisions and presents us with a bunch of questions like - What can we say about the success of a movie before it is released? Does the release day of the week have anything to do with the popularity and profit of the movie? Does the running time of a movie have an effect on its popularity and profitability of a movie? Can a probable revenue be predicted based on the pre-release data of a movie?

Objective

In this project, we'll be focussing on the following questions:

- Is there a relationship between the release day of the week and the movie's profitability and popularity? If yes, which weekdays as release days turn out to be most lucky for movies in terms of popularity and profit?
- How has the time duration been affecting high Profits, High Voting Average and High Popularity over the years from 2007 to 2017?
- Does the popularity of a director in a particular calendar year affect the chances of them being nominated for an award at the Cannes Film Festival?

In addition to this, if time permits, we will be modelling the movie revenue using machine learning techniques like regression, random forests, etc. using the pre-release data of the movie.

I have worked on the Movies Dataset:

The Movies Dataset

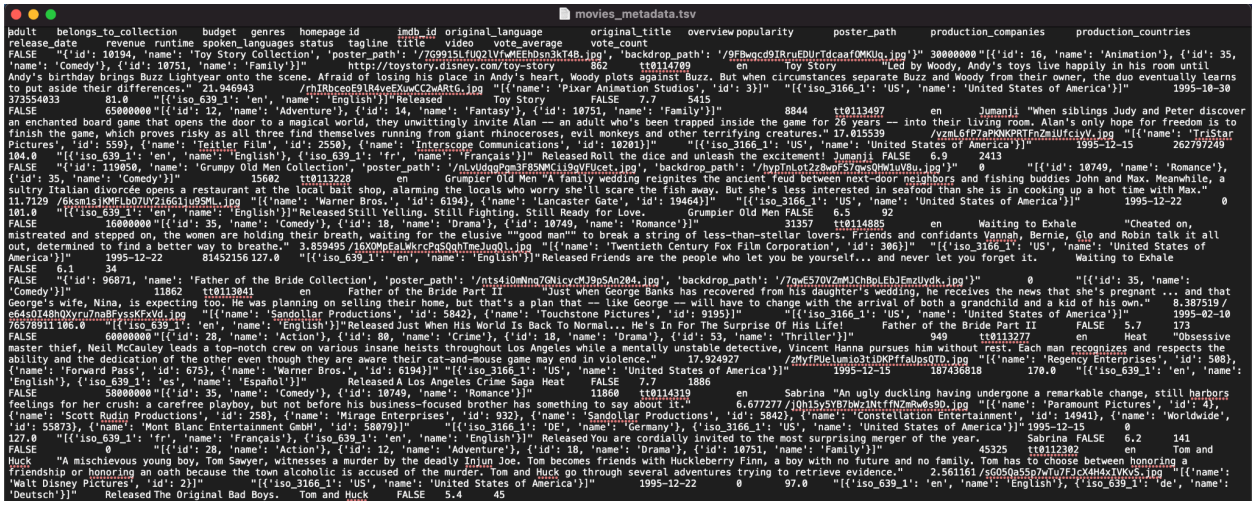
This dataset is an ensemble of data collected from TMDB and GroupLens. This is a database of 45000 movies released on or before July 2017. Data points include cast, crew, plot keywords, budget, revenue, posters, release dates, languages, production companies, countries, TMDB vote counts and vote averages. The files in this dataset that are useful to us are

- *movies_metadata.csv:* This has details about the genre, budget, language, popularity, etc.
- *credits.csv:* This contains the crew and cast information of the movies.

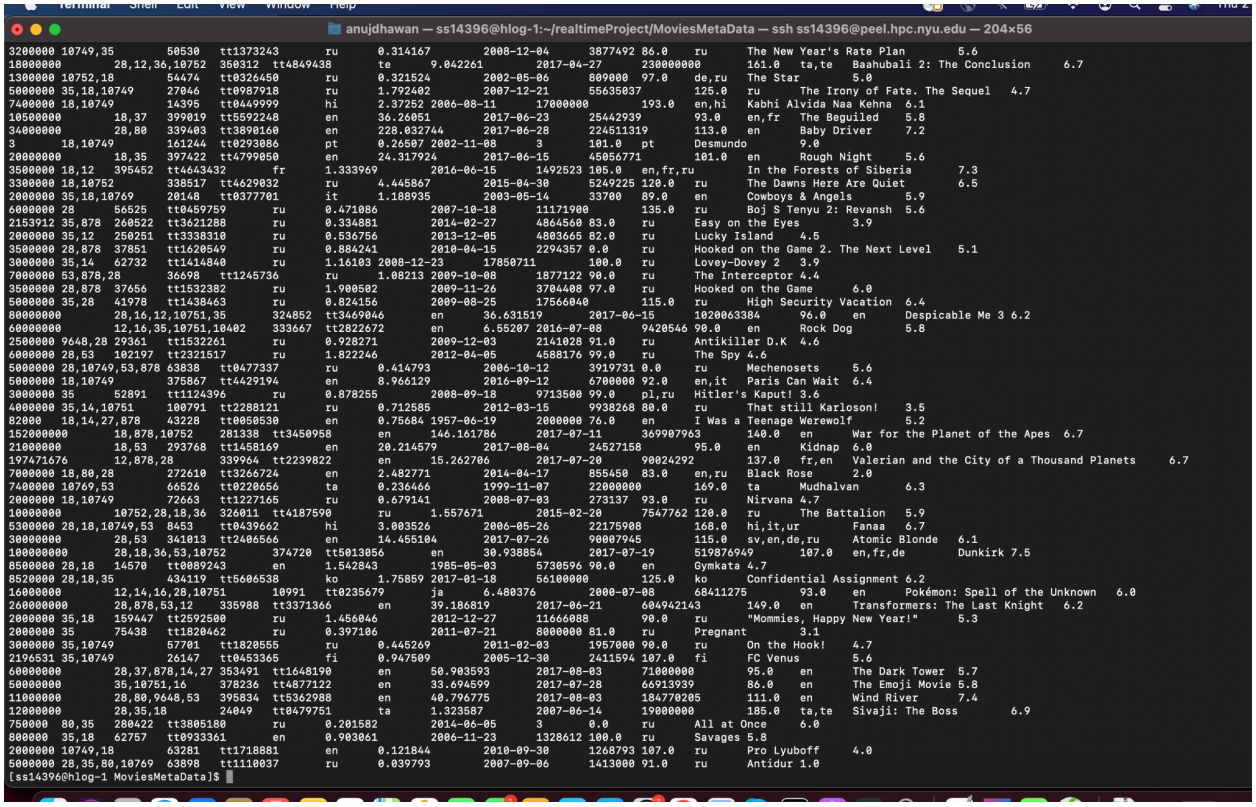
Data profiling, Cleaning, and Ingestion Tasks :

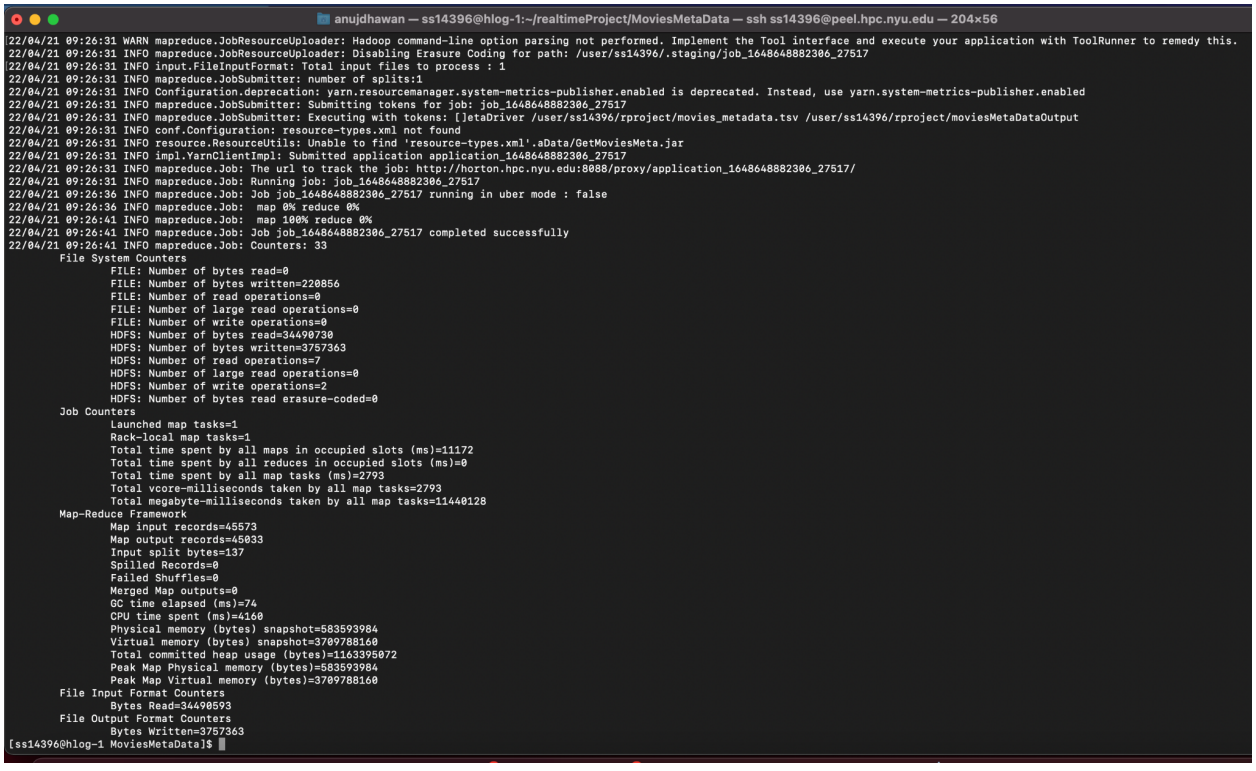
Movies_metadata.tsv :

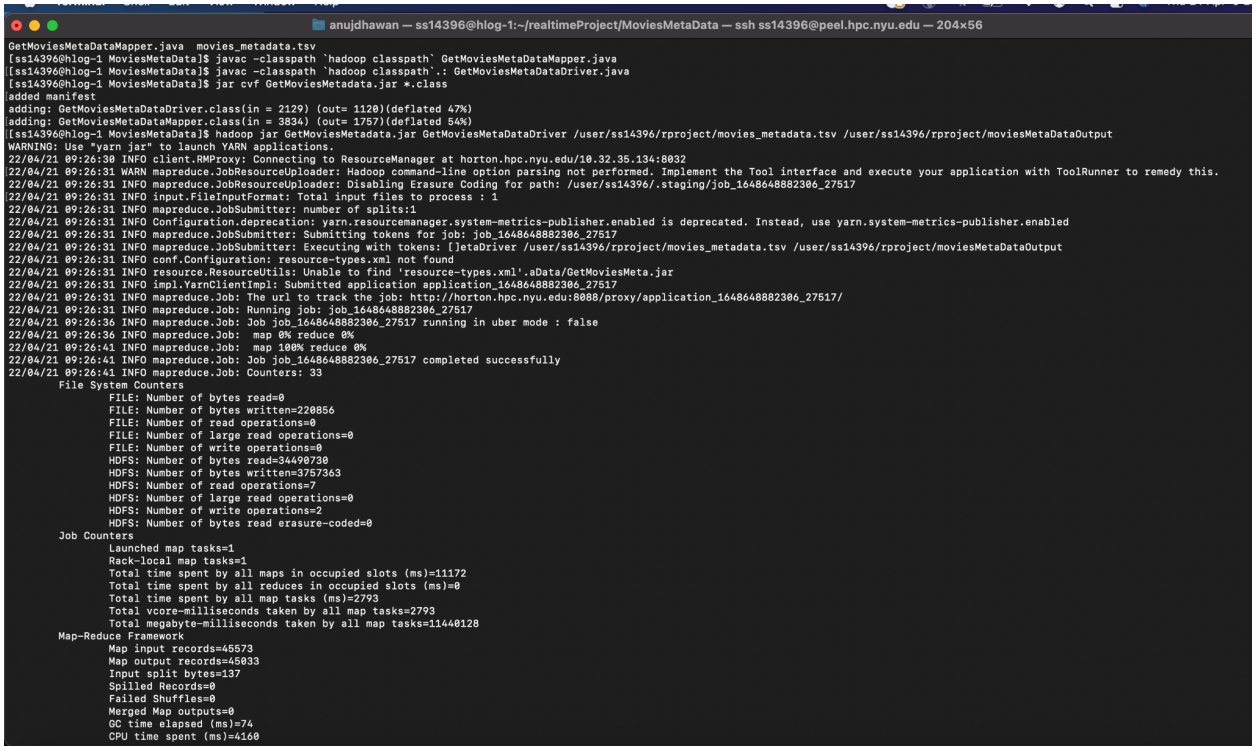
As the CSV file that JSON columns used tsv files so that tab could be used as a delimiter for extracting the rows. Attached is the snippet of the input :



- **For the movies dataset**, I removed the bad columns
  - That had NULL values
  - That had zero budget and revenue columns
  - That had empty JSON values
- The movies dataset had 24 columns, but only 10 columns were useful. Hence extracted those columns and wrote the final extracted columns to the output file.
- Attached a snippet of the output dataset and the shell commands for running the JAR file


anujdhawan — ss14396@hlog-1:~/realtimeProject/MoviesMetaData — ssh ss14396@peel.hpc.nyu.edu — 204×56
3200000 10749,35 50530 tt1373243 ru 0.314167 2008-12-04 3877492 86.0 ru The New Year's Rate Plan 5.6
18000000 28,12,36,10752 350312 tt4849438 te 9.042261 2017-04-27 230000000 161.0 ta,te Baahubali 2: The Conclusion 6.7
1300000 10752,18 54474 tt0326450 ru 0.321524 2002-05-06 809000 97.0 de,ru The Star 5.0
5000000 35,18,10749 27046 tt0987918 ru 1.792402 2007-12-21 55635037 125.0 ru The Irony of Fate. The Sequel 4.7
7400000 18,10749 14395 tt0449999 hi 2.37252 2006-08-11 17000000 193.0 en,hi Kabhi Alvida Naa Kehna 6.1
10500000 18,37 399019 tt5592248 en 36.26051 2017-06-23 25442939 93.0 en,fr The Beguiled 5.8
34000000 28,80 339403 tt3890160 en 228.032744 2017-06-28 224511319 113.0 en Baby Driver 7.2
3 18,10749 161244 tt0293086 pt 0.26507 2002-11-08 3 101.0 pt Desmundo 9.0
20000000 18,35 397422 tt4799050 en 24.317924 2017-06-15 45056771 101.0 en Rough Night 5.6
3500000 18,12 395452 tt4643432 fr 1.333969 2016-06-15 1492523 105.0 en,fr,ru In the Forests of Siberia 7.3
3300000 18,10752 338517 tt4629032 ru 4.445867 2015-04-30 5249225 120.0 ru The Dawns Here Are Quiet 6.5
2000000 35,18,10769 20148 tt0377701 it 1.188935 2003-05-14 33700 89.0 en Cowboys & Angels 5.9
6000000 28 56525 tt0459759 ru 0.471086 2007-10-18 11171900 135.0 ru Boj S Tenyu 2: Revansh 5.6
2153912 35,878 260522 tt3621288 ru 0.334881 2014-02-27 4864560 83.0 ru Easy on the Eyes 3.9
2000000 35,12 250251 tt3338310 ru 0.536756 2013-12-05 4803665 82.0 ru Lucky Island 4.5
3500000 28,878 37851 tt1620549 ru 0.884241 2010-04-15 2294357 0.0 ru Hooked on the Game 2. The Next Level 5.1
3000000 35,14 62732 tt1414840 ru 1.16103 2008-12-23 17850711 100.0 ru Lovey-Dovey 2 3.9
7000000 53,878,28 36698 tt1245736 ru 1.08213 2009-10-08 1877122 90.0 ru The Interceptor 4.4
3500000 28,878 37656 tt1532382 ru 1.900502 2009-11-26 3704408 97.0 ru Hooked on the Game 6.0
5000000 35,28 41978 tt1438463 ru 0.824156 2009-08-25 17566040 115.0 ru High Security Vacation 6.4
80000000 28,16,12,10751,35 324852 tt3469046 en 36.631519 2017-06-15 1020063384 96.0 en Despicable Me 3 6.2
60000000 12,16,35,10751,10402 333667 tt2822672 en 6.55207 2016-07-08 9420546 90.0 en Rock Dog 5.8
2500000 9648,28 29361 tt1532261 ru 0.928271 2009-12-03 2141028 91.0 ru Antikiller D.K 4.6
6000000 28,53 102197 tt2321517 ru 1.822246 2012-04-05 4588176 99.0 ru The Spy 4.6
5000000 28,10749,53,878 63838 tt0477337 ru 0.414793 2006-10-12 3919731 0.0 ru Mechenosets 5.6
5000000 18,10749 375867 tt4429194 en 8.966129 2016-09-12 6700000 92.0 en,it Paris Can Wait 6.4
3000000 35 52891 tt1124396 ru 0.878255 2008-09-18 9713500 99.0 pl,ru Hitler's Kaput! 3.6
4000000 35,14,10751 100791 tt2288121 ru 0.712585 2012-03-15 9938268 80.0 ru That still Karlson! 3.5
82000 18,14,27,878 43228 tt0050530 en 0.75684 1957-06-19 2000000 76.0 en I Was a Teenage Werewolf 5.2
152000000 18,878,10752 281338 tt3450958 en 146.161786 2017-07-11 369907963 140.0 en War for the Planet of the Apes 6.7
21000000 18,53 293768 tt1458169 en 20.214579 2017-08-04 24527158 95.0 en Kidnap 6.0
197471676 12,878,28 339964 tt2239822 en 15.262706 2017-07-20 90024292 137.0 fr,en Valerian and the City of a Thousand Planets 6.7
7000000 18,80,28 272610 tt3266724 en 2.482771 2014-04-17 855450 83.0 en,ru Black Rose 2.0
7400000 10769,53 66526 tt0220656 ta 0.236466 1999-11-07 22000000 169.0 ta Mudhalvan 6.3
2000000 18,10749 72663 tt1227165 ru 0.679141 2008-07-03 273137 93.0 ru Nirvana 4.7
10000000 10752,28,18,36 326011 tt4187590 ru 1.557671 2015-02-20 7547762 120.0 ru The Battalion 5.9
5300000 28,18,10749,53 8453 tt0439662 hi 3.003526 2006-05-26 22175908 168.0 hi,it,ur Fanaa 6.7
30000000 28,53 341013 tt2406566 en 14.455104 2017-07-26 90007945 115.0 sv,en,de,ru Atomic Blonde 6.1
100000000 28,18,36,53,10752 374720 tt5013056 en 30.938854 2017-07-19 519876949 107.0 en,fr,de Dunkirk 7.5
8500000 28,18 14570 tt0089243 en 1.542843 1985-05-03 5730596 90.0 en Gymkata 4.7
8520000 28,18,35 434119 tt5606538 ko 1.75859 2017-01-18 56100000 125.0 ko Confidential Assignment 6.2
16000000 12,14,16,28,10751 10991 tt0235679 ja 6.480376 2000-07-08 68411275 93.0 en Pokémon: Spell of the Unknown 6.0
260000000 28,878,53,12 335988 tt3371366 en 39.186819 2017-06-21 604942143 149.0 en Transformers: The Last Knight 6.2
2000000 35,18 159447 tt2592500 ru 1.456046 2012-12-27 11666088 90.0 ru "Mommies, Happy New Year!" 5.3
2000000 35 75438 tt1820462 ru 0.397106 2011-07-21 8000000 81.0 ru Pregnant 3.1
3000000 35,10749 57701 tt1820555 ru 0.445269 2011-02-03 1957000 90.0 ru On the Hook! 4.7
2196531 35,10749 26147 tt0453365 fi 0.947509 2005-12-30 2411594 107.0 fi FC Venus 5.6
60000000 28,37,878,14,27 353491 tt1648190 en 50.903593 2017-08-03 71000000 95.0 en The Dark Tower 5.7
50000000 35,10751,16 378236 tt4877122 en 33.694599 2017-07-28 66913939 86.0 en The Emoji Movie 5.8
11000000 28,80,9648,53 395834 tt5362988 en 40.796775 2017-08-03 184770205 111.0 en Wind River 7.4
12000000 28,35,18 24049 tt0479751 ta 1.323587 2007-06-14 19000000 185.0 ta,te Sivaji: The Boss 6.9
750000 80,35 280422 tt3805180 ru 0.201582 2014-06-05 3 0.0 ru All at Once 6.0
800000 35,18 62757 tt0933361 en 0.903061 2006-11-23 1328612 100.0 ru Savages 5.8
2000000 10749,18 63281 tt1718881 en 0.121844 2010-09-30 1268793 107.0 ru Pro Lyuboff 4.0
5000000 28,35,80,10769 63898 tt1110037 ru 0.039793 2007-09-06 1413000 91.0 ru Antidur 1.0
[ss14396@hlog-1 MoviesMetaData]$


anujdhawan — ss14396@hlog-1:~/realtimeProject/MoviesMetaData — ssh ss14396@peel.hpc.nyu.edu — 204×56
22/04/21 09:26:31 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
22/04/21 09:26:31 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /user/ss14396/.staging/job_1648648882306_27517
22/04/21 09:26:31 INFO input.FileInputFormat: Total input files to process : 1
22/04/21 09:26:31 INFO mapreduce.JobSubmitter: number of splits:1
22/04/21 09:26:31 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
22/04/21 09:26:31 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1648648882306_27517
22/04/21 09:26:31 INFO mapreduce.JobSubmitter: Executing with tokens: []etaDriver /user/ss14396/rproject/movies_metadata.tsv /user/ss14396/rproject/moviesMetaDataOutput
22/04/21 09:26:31 INFO conf.Configuration: resource-types.xml not found
22/04/21 09:26:31 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.aData/GetMoviesMeta.jar
22/04/21 09:26:31 INFO impl.YarnClientImpl: Submitted application application_1648648882306_27517
22/04/21 09:26:31 INFO mapreduce.Job: The url to track the job: http://horton.hpc.nyu.edu:8088/proxy/application_1648648882306_27517/
22/04/21 09:26:31 INFO mapreduce.Job: Running job: job_1648648882306_27517
22/04/21 09:26:36 INFO mapreduce.Job: Job job_1648648882306_27517 running in uber mode : false
22/04/21 09:26:36 INFO mapreduce.Job: map 0% reduce 0%
22/04/21 09:26:41 INFO mapreduce.Job: map 100% reduce 0%
22/04/21 09:26:41 INFO mapreduce.Job: Job job_1648648882306_27517 completed successfully
22/04/21 09:26:41 INFO mapreduce.Job: Counters: 33
File System Counters
FILE: Number of bytes read=0
FILE: Number of bytes written=220856
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=34490730
HDFS: Number of bytes written=3757363
HDFS: Number of read operations=7
HDFS: Number of large read operations=0
HDFS: Number of write operations=2
HDFS: Number of bytes read erasure-coded=0
Job Counters
Launched map tasks=1
Rack-local map tasks=1
Total time spent by all maps in occupied slots (ms)=11172
Total time spent by all reduces in occupied slots (ms)=0
Total time spent by all map tasks (ms)=2793
Total vcore-milliseconds taken by all map tasks=2793
Total megabyte-milliseconds taken by all map tasks=11440128
Map-Reduce Framework
Map input records=45573
Map output records=45033
Input split bytes=137
Spilled Records=0
Failed Shuffles=0
Merged Map outputs=0
GC time elapsed (ms)=74
CPU time spent (ms)=4160
Physical memory (bytes) snapshot=583593984
Virtual memory (bytes) snapshot=3709788160
Total committed heap usage (bytes)=1163395072
Peak Map Physical memory (bytes)=583593984
Peak Map Virtual memory (bytes)=3709788160
File Input Format Counters
Bytes Read=34490593
File Output Format Counters
Bytes Written=3757363
[ss14396@hlog-1 MoviesMetaData]$



Credits.tsv :

As the CSV file that JSON columns used tsv files so that tab could be used as a delimiter for extracting the rows. Attached is the snippet of the input :

For the Credits.tsv 4 tasks needed to be done :

- **Create a mapping between CAST ID and cast name**
- **Create a mapping between director ID and director name**

- **Create a mapping between the cast and the list of movies it has been on**
- **Create a mapping between the movies and the list of cast**

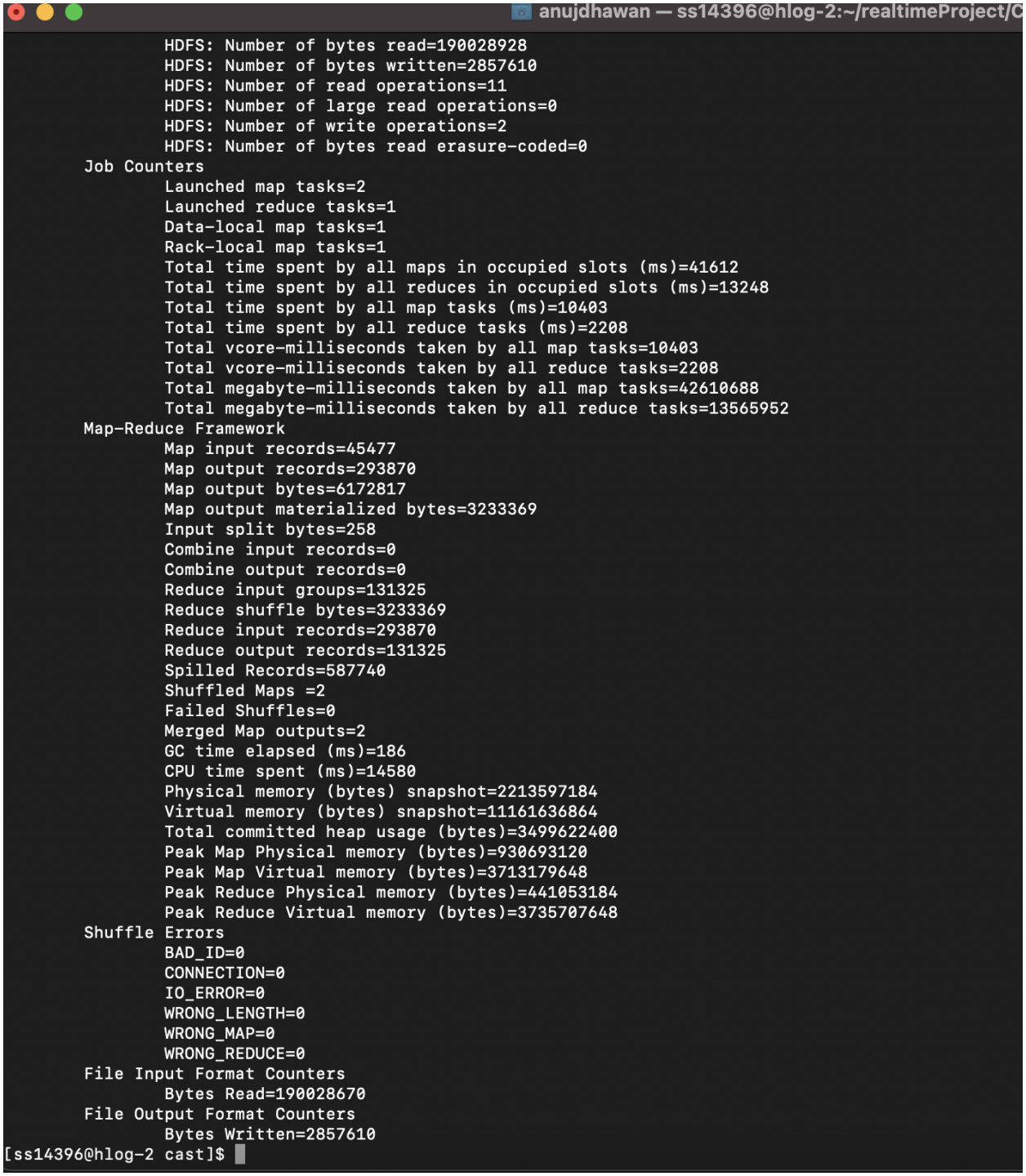
All the code and jar files are present in the credits folder.

Create a mapping between CAST ID and cast name :

- Firstly, removed the bad columns
    - That had NULL values
    - That had zero budget and revenue columns
    - That had empty JSON values

- For this, from the cast column of the credits.tsv dataset we extract the name and id of every cast and create a mapping from
- All the code and JAR files are present in cast folder
- Attached a snippet of the output dataset and the shell commands for running the JAR file


anujdhawan — ss14396@hlog-2
999575 James Kautz
999576 Courtney-Anne Doody
999579 Brandon Boyd
99958 Sadou Teymouri
99959 Hoyatala Hakimi
999598 Patricia Gozzi
9996 Sean Whalen
999620 Aleksandr Medvedkin
999621 Pyotr Zinovyev
999622 Yelena Yegorova
999623 Mikhail Gipsi
99963 Gordon Oas-Heim
999630 Niels Nørløv Hansen
999636 Kari Hevossaari
999637 Juha Hippi
999672 Michael Adams
99968 Ben Moore
9997 Scott Thomson
999700 Rock A. Walker
999716 Gary Combs
999718 Cicely Courtneidge
999725 Masaaki Uchino
999729 Alexey Bardukov
99973 Shelby Livingston
999734 Alexey Morozov
999736 Bianca Hunter
999757 Engin Alpateş
999758 Tansu Bicer
999759 Köksal Engür
99978 Candi Conder
999790 Alexis Knapp
9998 Joey Slotnick
999800 Arnaud Cosson
999803 Antoine Levannier
999804 Frédéric Kontogom
999817 Olivia Taylor Dudley
99982 Elyn Warner
999820 Monica Clay
999822 Tyler Maynard
999838 Diego Torres
999842 José Luis García
999843 Joaquín Cascales Zarzo
999844 Alberto Jiménez
999860 David Osterhout
999870 Tiago Correa
999879 Jeremy Clark
999880 Holly Lynn Ellis
999881 Garth Blomberg
999883 Brent Mydland
999904 Shannah Laumeister
999939 Joyce Bland
99994 Michael Gingold
99996 Scooter McCrae
999996 María Douglas
999999 Osvaldo Bonet
[ss14396@hlog-2 cast]$



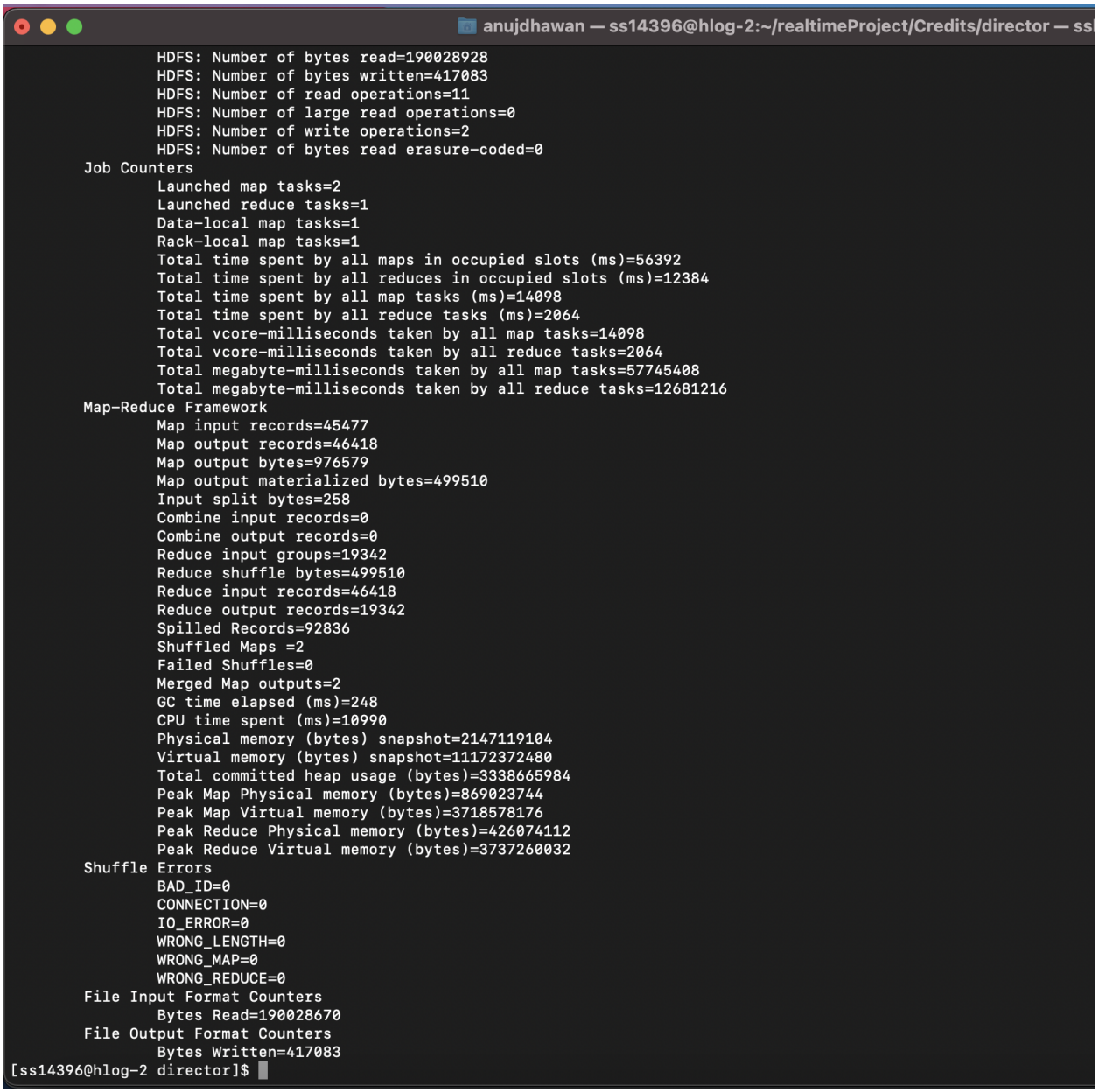
Create a mapping between Director ID and director name :

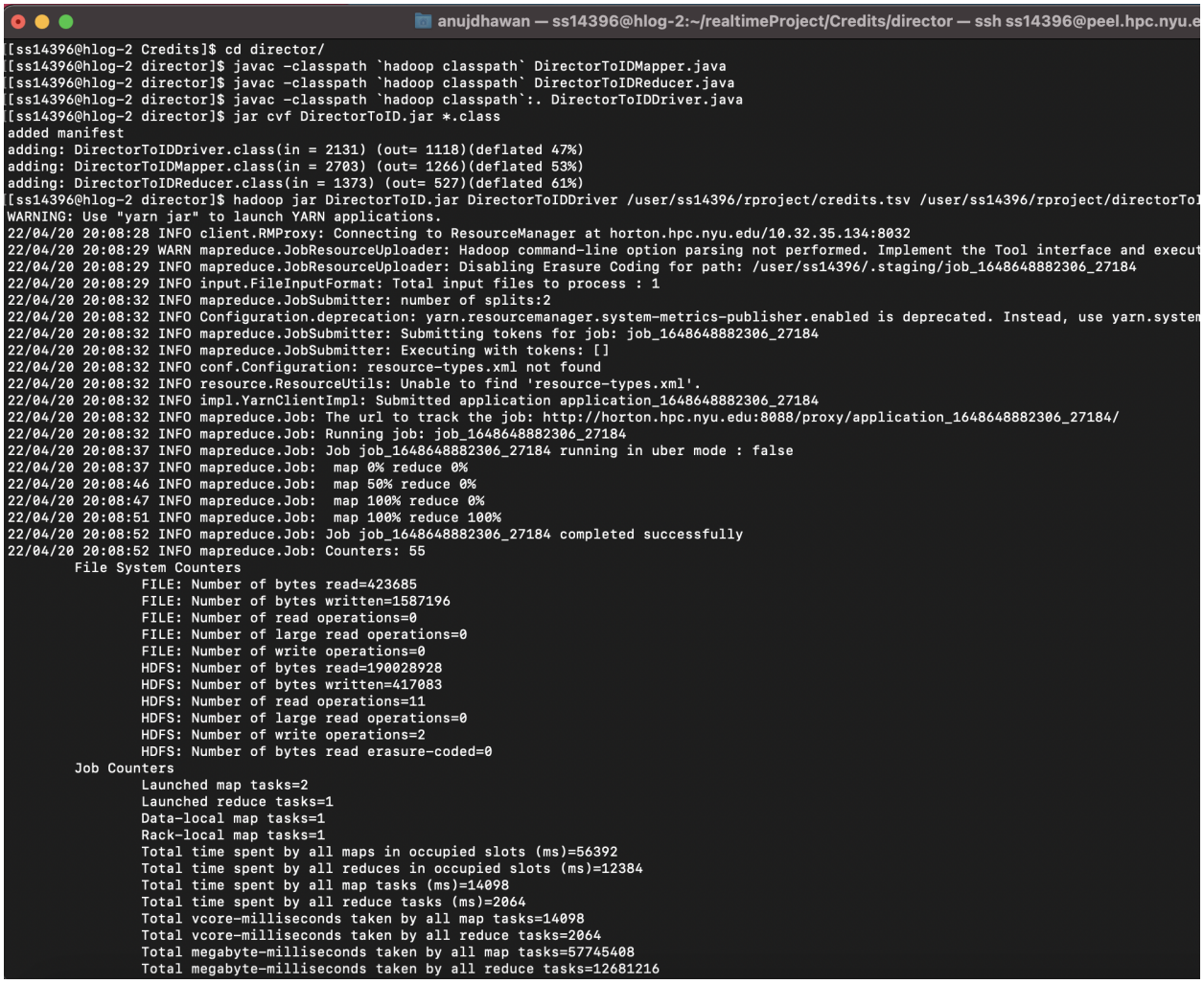
- Firstly, removed the bad columns

    - That had NULL values
    - That had zero budget and revenue columns
    - That had empty JSON values

- For this, from the crew column of the credits.tsv dataset we extract the name and id of the crew that has the job of director and create a mapping from
- All the code and JAR files are present in director folder
- Attached a snippet of the output dataset and the shell commands for running the JAR file


anujdhawan — ss14396@hlog-2:~/real
99577 Vernon Sewell
996070 Charles Adelman
996242 Tyler Spindel
99627 Michael Paul Girard
99651 Yan Frid
996539 David Lee Miller
99667 Alexander Hall
996672 Przemyslaw Angerman
996681 Tyron Montgomery
996684 Tim Coleman
996685 Boris Zubov
996749 Zoe Clarke-Williams
996817 Vicente Franco
99684 Corbin Timbrook
99689 Deng Chao
99710 Vidhu Vinod Chopra
997235 Jeff Newman
99747 Jesper W. Nielsen
997560 Brett Winn
997630 Vanessa Hope
997694 Germaine Dulac
997700 Ellen Spiro
997819 Deepak Sareen
99805 Clay Westervelt
998227 Thomas Zellen
99836 John Peyser
998473 Strathford Hamilton
998507 Stig Svendsen
998511 Lane Janger
998534 Tomu Uchida
998555 Victor Kossakovsky
998589 John Swanbeck
99859 Tim McCann
998614 Viking Eggeling
998745 Carlos Marcovich
99875 Masami Hata
99884 Lawrence Huntington
998842 Efram Potelle
99898 David Chase
99904 Robert Oppel
99916 Herschell Gordon Lewis
9994 Helen Hunt
999545 Dave Edwards
999558 David Mason
999559 Debbi Slater
999620 Aleksandr Medvedkin
999630 Niels Nørløv Hansen
999717 Mark Edlitz
999723 Nobuo Mizuta
999760 Eran Creevy
99979 Servando González
999845 Florián Rey
999882 Nirpal Bhogal
999909 Dusty Bias
99996 Scooter McCrae
[ss14396@hlog-2 director]$

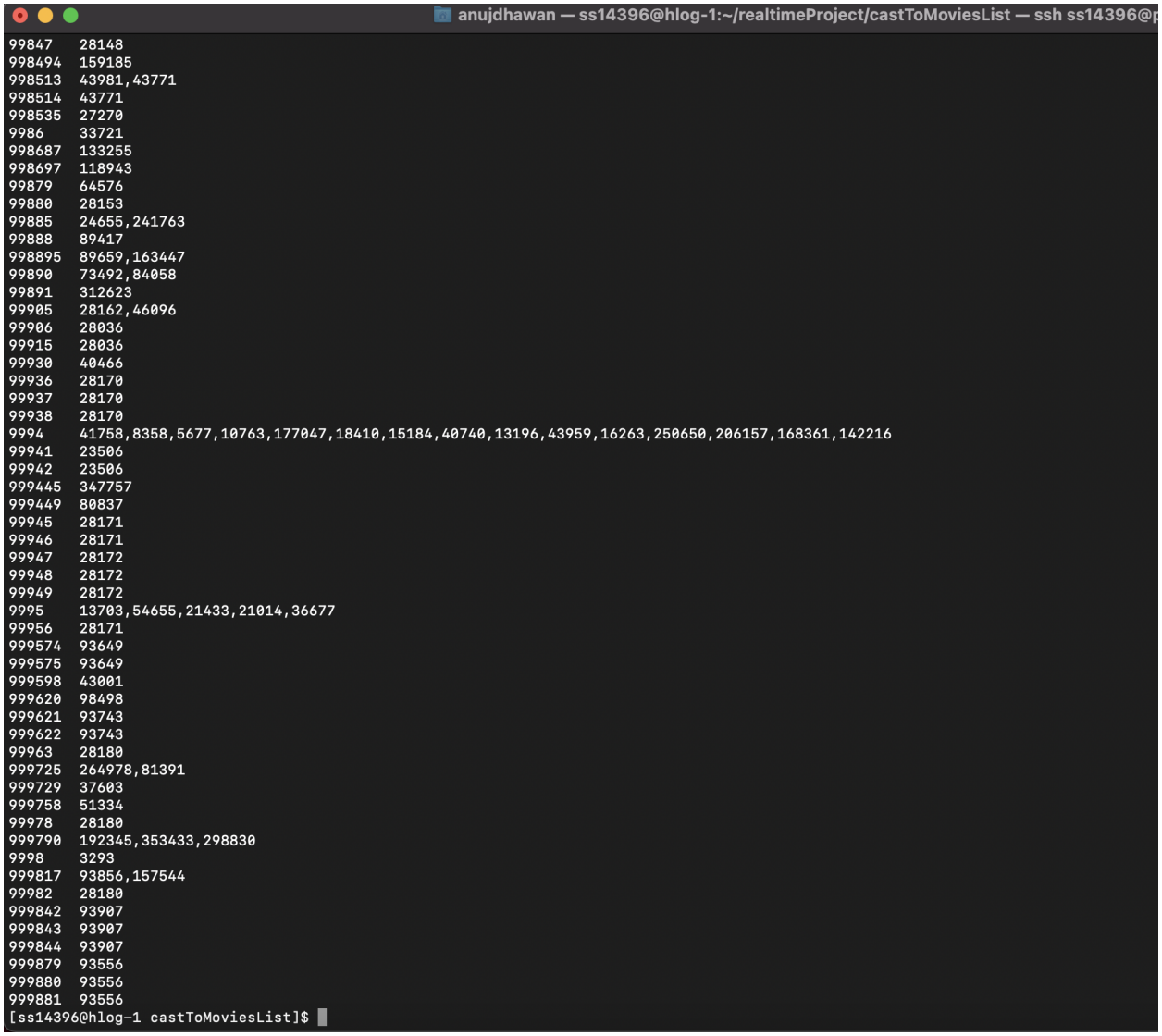

anujdhawan — ss14396@hlog-2:~/realtimeProject/Credits/director — ss
HDFS: Number of bytes read=190028928
HDFS: Number of bytes written=417083
HDFS: Number of read operations=11
HDFS: Number of large read operations=0
HDFS: Number of write operations=2
HDFS: Number of bytes read erasure-coded=0
Job Counters
Launched map tasks=2
Launched reduce tasks=1
Data-local map tasks=1
Rack-local map tasks=1
Total time spent by all maps in occupied slots (ms)=56392
Total time spent by all reduces in occupied slots (ms)=12384
Total time spent by all map tasks (ms)=14098
Total time spent by all reduce tasks (ms)=2064
Total vcore-milliseconds taken by all map tasks=14098
Total vcore-milliseconds taken by all reduce tasks=2064
Total megabyte-milliseconds taken by all map tasks=57745408
Total megabyte-milliseconds taken by all reduce tasks=12681216
Map-Reduce Framework
Map input records=45477
Map output records=46418
Map output bytes=976579
Map output materialized bytes=499510
Input split bytes=258
Combine input records=0
Combine output records=0
Reduce input groups=19342
Reduce shuffle bytes=499510
Reduce input records=46418
Reduce output records=19342
Spilled Records=92836
Shuffled Maps =2
Failed Shuffles=0
Merged Map outputs=2
GC time elapsed (ms)=248
CPU time spent (ms)=10990
Physical memory (bytes) snapshot=2147119104
Virtual memory (bytes) snapshot=11172372480
Total committed heap usage (bytes)=3338665984
Peak Map Physical memory (bytes)=869023744
Peak Map Virtual memory (bytes)=3718578176
Peak Reduce Physical memory (bytes)=426074112
Peak Reduce Virtual memory (bytes)=3737260032
Shuffle Errors
BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0
File Input Format Counters
Bytes Read=190028670
File Output Format Counters
Bytes Written=417083
[ss14396@hlog-2 director]$

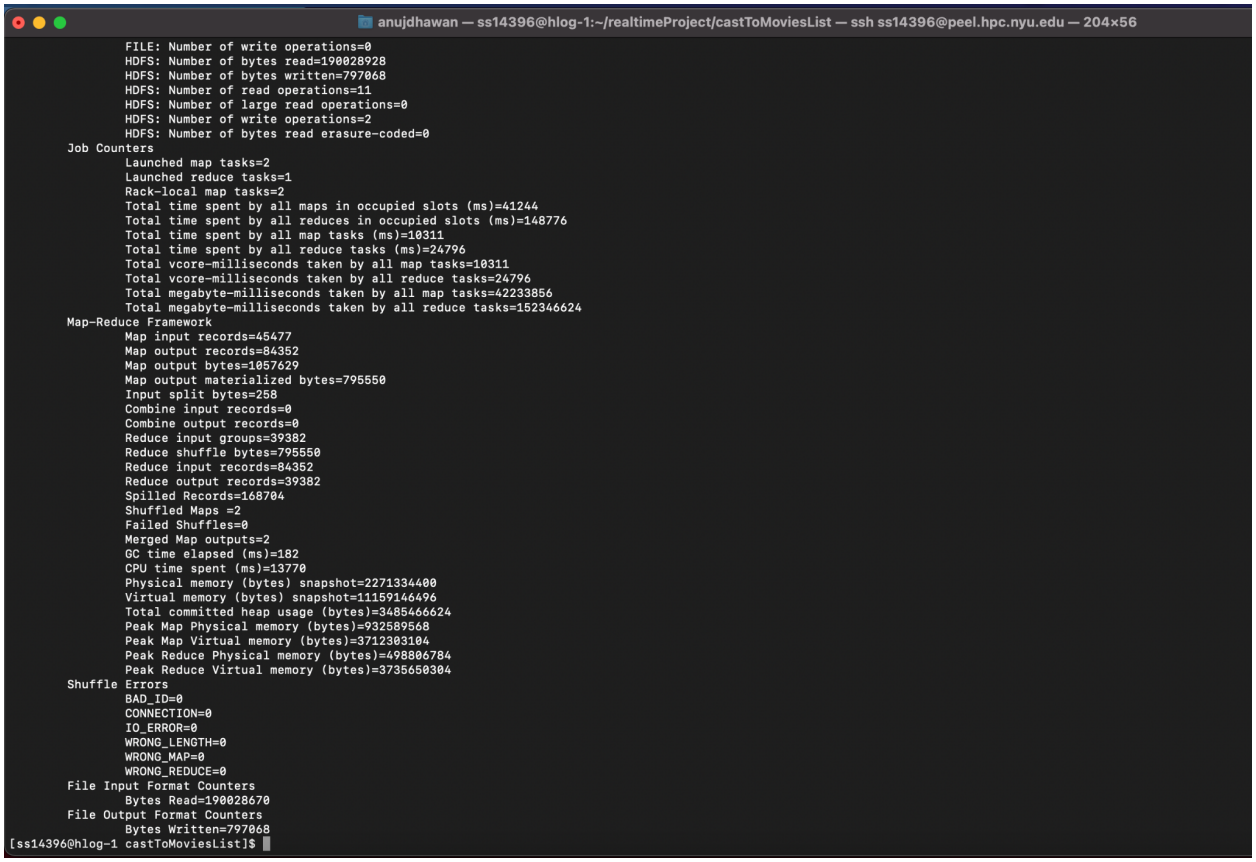


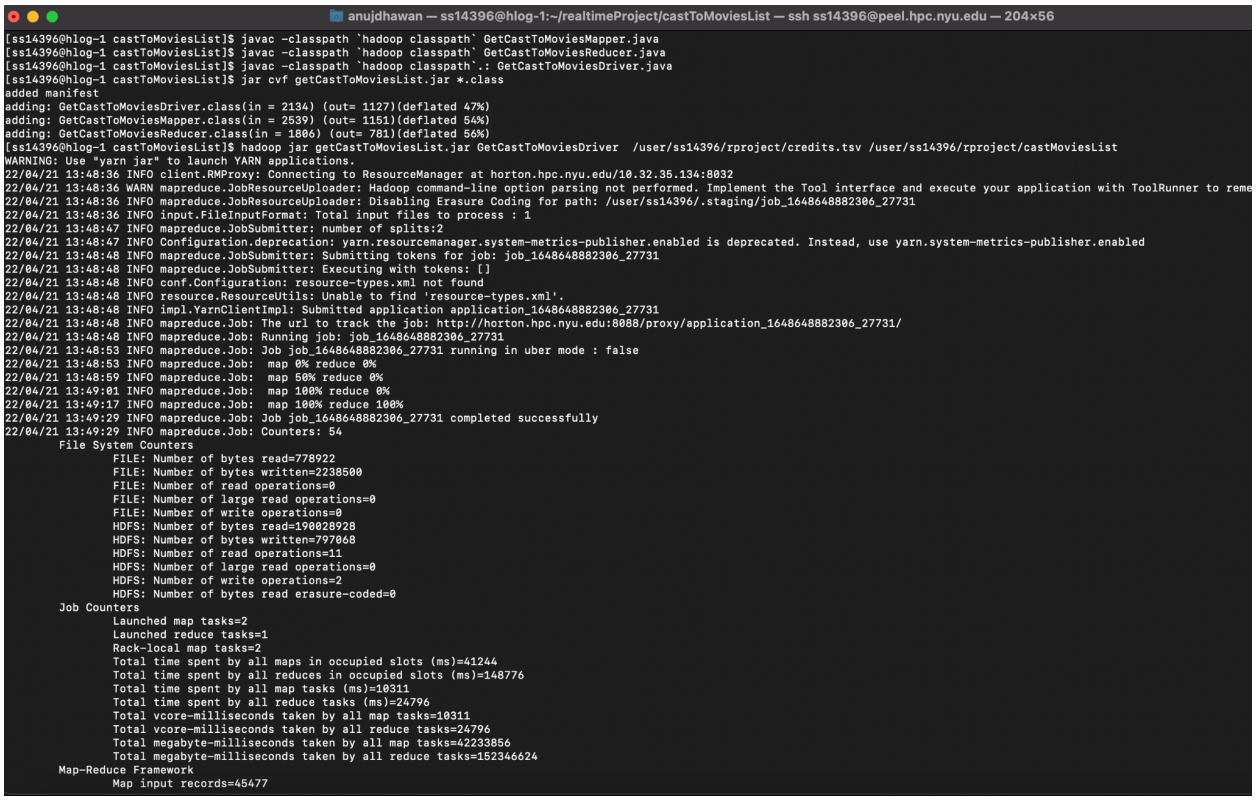
Create a mapping between the cast and the list of movies it has been on

- Firstly, removed the bad columns
    - That had NULL values
    - That had zero budget and revenue columns
    - That had empty JSON values

- For this, from the cast column of the credits.tsv dataset we extract the id of the cast as key and movie ID as the value in the mapper task. Then all the movie ids for a particular cast id are combined in the reducer task
- All the code and JAR files are present in the castToMoviesList folder
- We have only considered the top 3 casts. JSON column had cast in decreasing order of their importance.
- Attached a snippet of the output dataset and the shell commands for running the JAR file


anujdhawan — ss14396@hlog-1:~/realtimeProject/castToMoviesList — ssh ss14396@p
99847 28148
998494 159185
998513 43981,43771
998514 43771
998535 27270
9986 33721
998687 133255
998697 118943
99879 64576
99880 28153
99885 24655,241763
99888 89417
998895 89659,163447
99890 73492,84058
99891 312623
99905 28162,46096
99906 28036
99915 28036
99930 40466
99936 28170
99937 28170
99938 28170
9994 41758,8358,5677,10763,177047,18410,15184,40740,13196,43959,16263,250650,206157,168361,142216
99941 23506
99942 23506
999445 347757
999449 80837
99945 28171
99946 28171
99947 28172
99948 28172
99949 28172
9995 13703,54655,21433,21014,36677
99956 28171
999574 93649
999575 93649
999598 43001
999620 98498
999621 93743
999622 93743
99963 28180
999725 264978,81391
999729 37603
999758 51334
99978 28180
999790 192345,353433,298830
9998 3293
999817 93856,157544
99982 28180
999842 93907
999843 93907
999844 93907
999879 93556
999880 93556
999881 93556
[ss14396@hlog-1 castToMoviesList]$

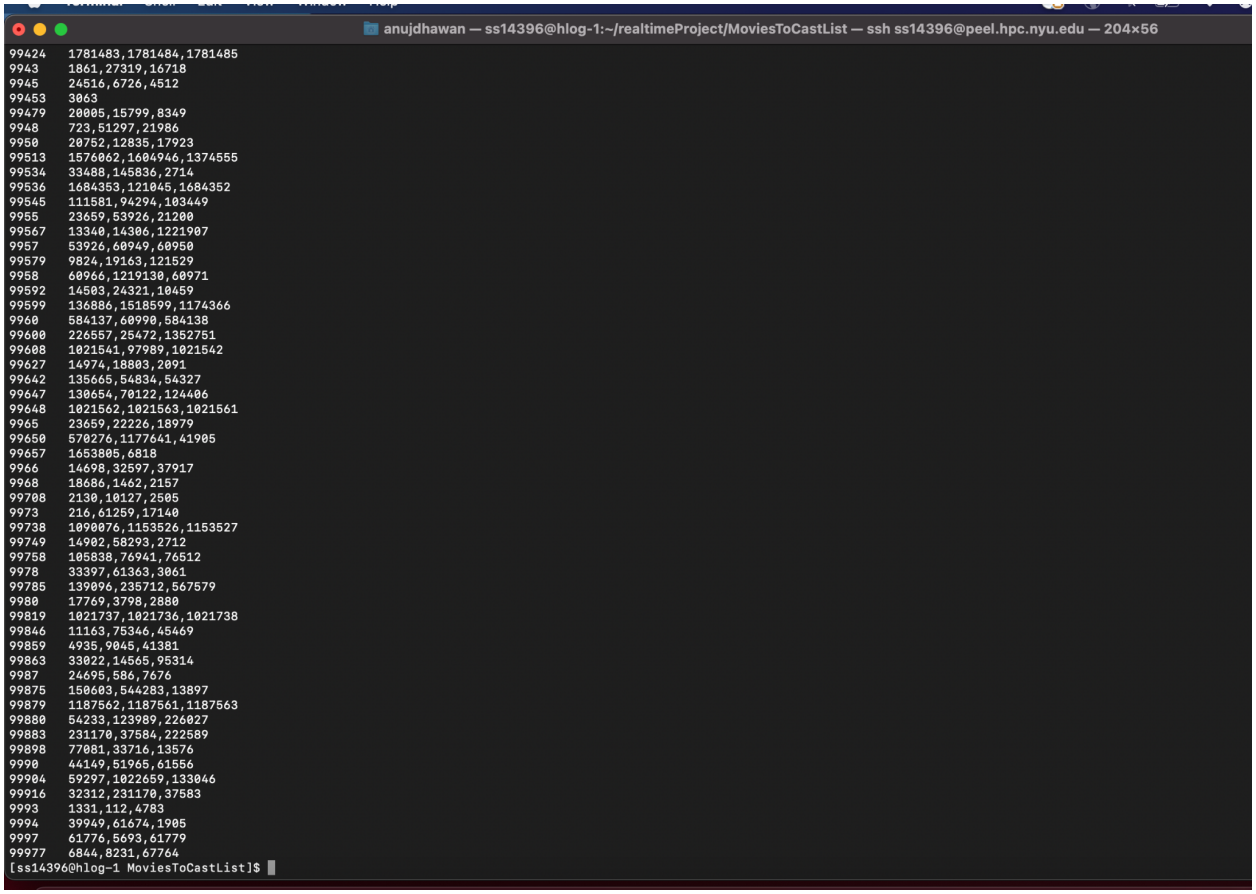

anujdhawan — ss14396@hlog-1:~/realtimeProject/castToMoviesList — ssh ss14396@peel.hpc.nyu.edu — 204×56
FILE: Number of write operations=0
HDFS: Number of bytes read=190028928
HDFS: Number of bytes written=797068
HDFS: Number of read operations=11
HDFS: Number of large read operations=0
HDFS: Number of write operations=2
HDFS: Number of bytes read erasure-coded=0
Job Counters
Launched map tasks=2
Launched reduce tasks=1
Rack-local map tasks=2
Total time spent by all maps in occupied slots (ms)=41244
Total time spent by all reduces in occupied slots (ms)=148776
Total time spent by all map tasks (ms)=10311
Total time spent by all reduce tasks (ms)=24796
Total vcore-milliseconds taken by all map tasks=10311
Total vcore-milliseconds taken by all reduce tasks=24796
Total megabyte-milliseconds taken by all map tasks=42233856
Total megabyte-milliseconds taken by all reduce tasks=152346624
Map-Reduce Framework
Map input records=45477
Map output records=84352
Map output bytes=1057629
Map output materialized bytes=795550
Input split bytes=258
Combine input records=0
Combine output records=0
Reduce input groups=39382
Reduce shuffle bytes=795550
Reduce input records=84352
Reduce output records=39382
Spilled Records=168704
Shuffled Maps =2
Failed Shuffles=0
Merged Map outputs=2
GC time elapsed (ms)=182
CPU time spent (ms)=13770
Physical memory (bytes) snapshot=2271334400
Virtual memory (bytes) snapshot=11159146496
Total committed heap usage (bytes)=3485466624
Peak Map Physical memory (bytes)=932589568
Peak Map Virtual memory (bytes)=3712303104
Peak Reduce Physical memory (bytes)=498806784
Peak Reduce Virtual memory (bytes)=3735650304
Shuffle Errors
BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0
File Input Format Counters
Bytes Read=190028670
File Output Format Counters
Bytes Written=797068
[ss14396@hlog-1 castToMoviesList]$

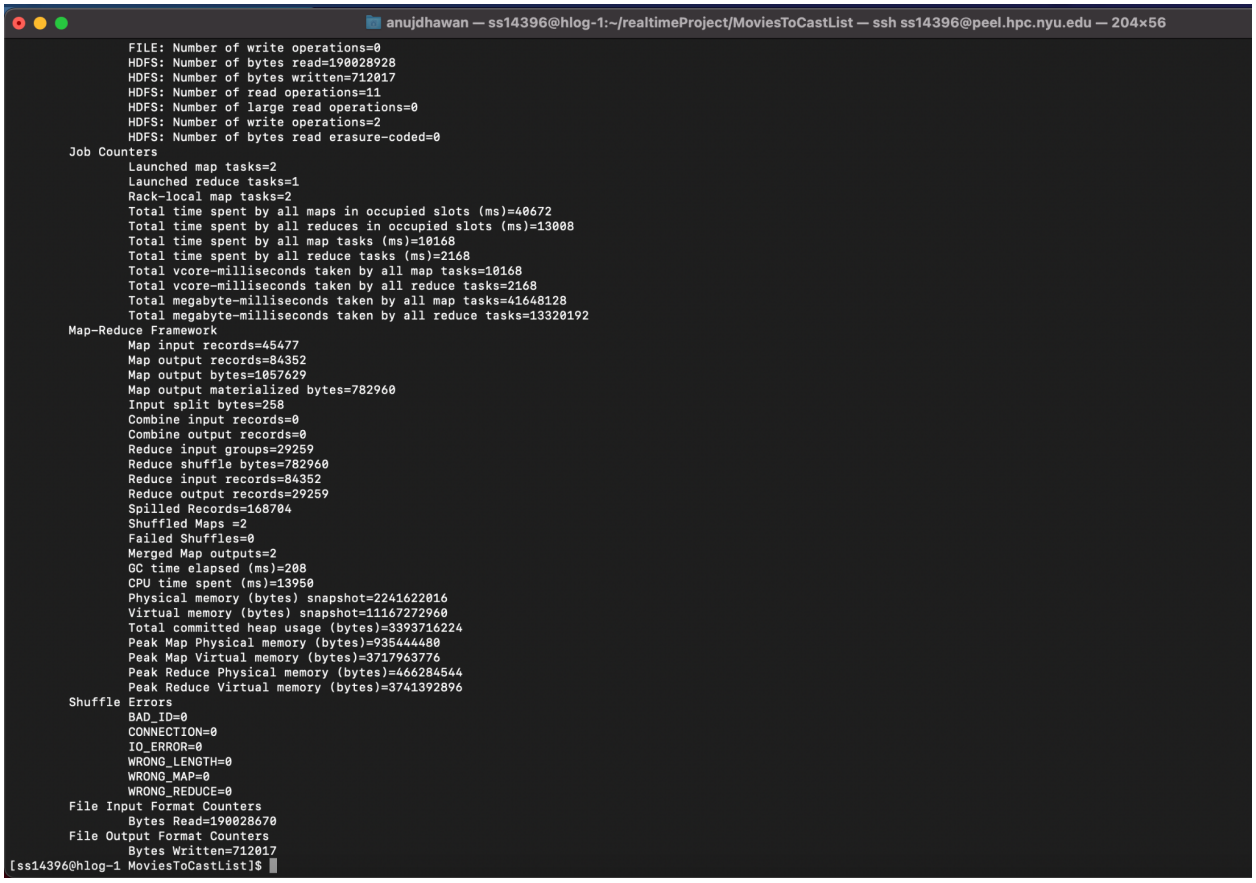

anujdhawan — ss14396@hlog-1:~/realtimeProject/castToMoviesList — ssh ss14396@peel.hpc.nyu.edu — 204×56
[ss14396@hlog-1 castToMoviesList]$ javac -classpath `hadoop classpath` GetCastToMoviesMapper.java
[ss14396@hlog-1 castToMoviesList]$ javac -classpath `hadoop classpath` GetCastToMoviesReducer.java
[ss14396@hlog-1 castToMoviesList]$ javac -classpath `hadoop classpath`.: GetCastToMoviesDriver.java
[ss14396@hlog-1 castToMoviesList]$ jar cvf getCastToMoviesList.jar *.class
added manifest
adding: GetCastToMoviesDriver.class(in = 2134) (out= 1127)(deflated 47%)
adding: GetCastToMoviesMapper.class(in = 2539) (out= 1151)(deflated 54%)
adding: GetCastToMoviesReducer.class(in = 1806) (out= 781)(deflated 56%)
[ss14396@hlog-1 castToMoviesList]$ hadoop jar getCastToMoviesList.jar GetCastToMoviesDriver /user/ss14396/rproject/credits.tsv /user/ss14396/rproject/castMoviesList
WARNING: Use "yarn jar" to launch YARN applications.
22/04/21 13:48:36 INFO client.RMProxy: Connecting to ResourceManager at horton.hpc.nyu.edu/10.32.35.134:8032
22/04/21 13:48:36 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to reme
22/04/21 13:48:36 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /user/ss14396/.staging/job_1648648882306_27731
22/04/21 13:48:36 INFO input.FileInputFormat: Total input files to process : 1
22/04/21 13:48:47 INFO mapreduce.JobSubmitter: number of splits:2
22/04/21 13:48:47 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
22/04/21 13:48:48 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1648648882306_27731
22/04/21 13:48:48 INFO mapreduce.JobSubmitter: Executing with tokens: []
22/04/21 13:48:48 INFO conf.Configuration: resource-types.xml not found
22/04/21 13:48:48 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
22/04/21 13:48:48 INFO impl.YarnClientImpl: Submitted application application_1648648882306_27731
22/04/21 13:48:48 INFO mapreduce.Job: The url to track the job: http://horton.hpc.nyu.edu:8088/proxy/application_1648648882306_27731/
22/04/21 13:48:48 INFO mapreduce.Job: Running job: job_1648648882306_27731
22/04/21 13:48:53 INFO mapreduce.Job: Job job_1648648882306_27731 running in uber mode : false
22/04/21 13:48:53 INFO mapreduce.Job: map 0% reduce 0%
22/04/21 13:48:59 INFO mapreduce.Job: map 50% reduce 0%
22/04/21 13:49:01 INFO mapreduce.Job: map 100% reduce 0%
22/04/21 13:49:17 INFO mapreduce.Job: map 100% reduce 100%
22/04/21 13:49:29 INFO mapreduce.Job: Job job_1648648882306_27731 completed successfully
22/04/21 13:49:29 INFO mapreduce.Job: Counters: 54
File System Counters
FILE: Number of bytes read=778922
FILE: Number of bytes written=2238500
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=190028928
HDFS: Number of bytes written=797068
HDFS: Number of read operations=11
HDFS: Number of large read operations=0
HDFS: Number of write operations=2
HDFS: Number of bytes read erasure-coded=0
Job Counters
Launched map tasks=2
Launched reduce tasks=1
Rack-local map tasks=2
Total time spent by all maps in occupied slots (ms)=41244
Total time spent by all reduces in occupied slots (ms)=148776
Total time spent by all map tasks (ms)=10311
Total time spent by all reduce tasks (ms)=24796
Total vcore-milliseconds taken by all map tasks=10311
Total vcore-milliseconds taken by all reduce tasks=24796
Total megabyte-milliseconds taken by all map tasks=42233856
Total megabyte-milliseconds taken by all reduce tasks=152346624
Map-Reduce Framework
Map input records=45477

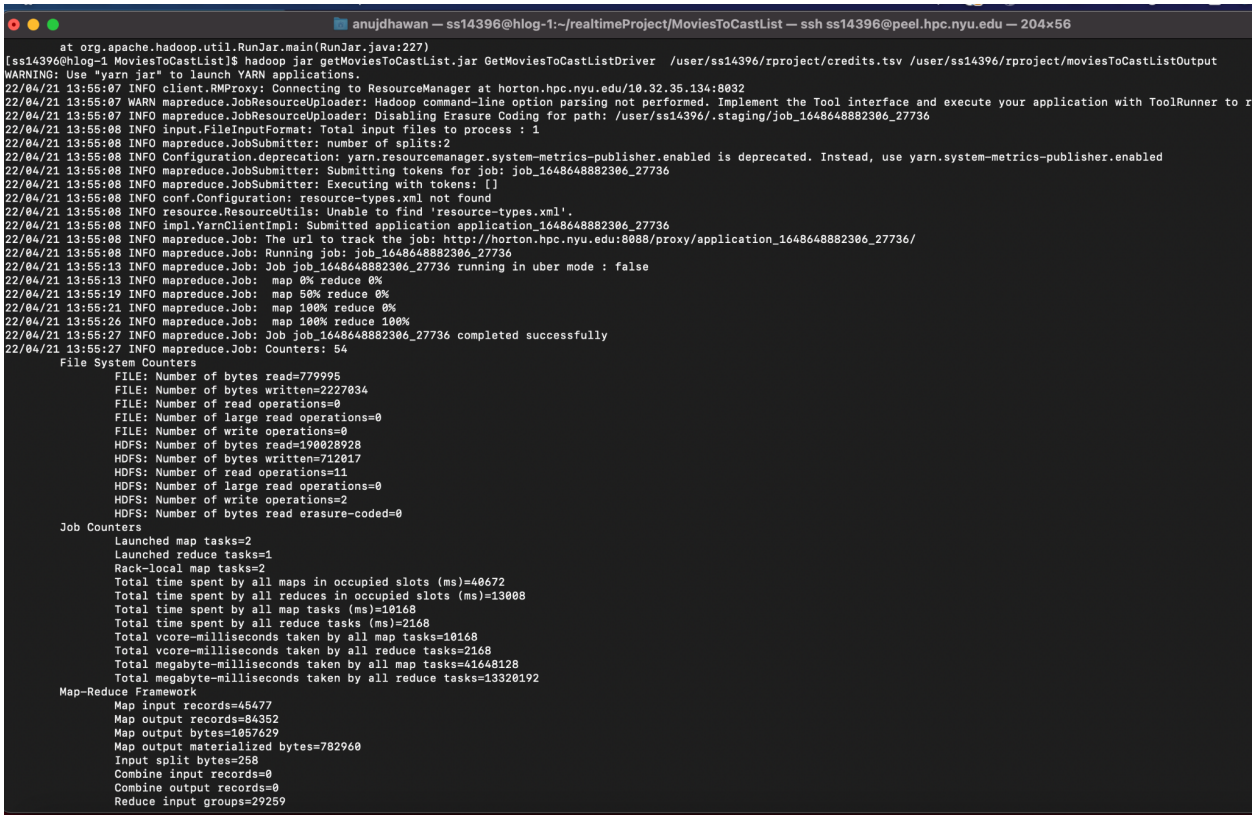
Create a mapping between the movie ID and the list of cast ID being featured in that

- Firstly, removed the bad columns
    - That had NULL values
    - That had zero budget and revenue columns
    - That had empty JSON values

- For this, from the cast column of the credits.tsv dataset we extract the id of the cast as the value with movie ID as the key in the mapper task. Then all the cast ids for a particular movie id are combined in the reducer task
- We have only considered the top 3 casts. JSON column had cast in decreasing order of their importance.
- All the code and JAR files are present in the moviesToCastList folder
- Attached is a snippet of the output dataset and the shell commands for running the JAR file


anujdhawan — ss14396@hlog-1:~/realtimeProject/MoviesToCastList — ssh ss14396@peel.hpc.nyu.edu — 204×56
99424 1781483,1781484,1781485
9943 1861,27319,16718
9945 24516,6726,4512
99453 3063
99479 20005,15799,8349
9948 723,51297,21986
9950 20752,12835,17923
99513 1576062,1604946,1374555
99534 33488,145836,2714
99536 1684353,121045,1684352
99545 111581,94294,103449
9955 23659,53926,21200
99567 13340,14306,1221907
9957 53926,60949,60950
99579 9824,19163,121529
9958 60966,1219130,60971
99592 14503,24321,10459
99599 136886,1518599,1174366
9960 584137,60990,584138
99600 226557,25472,1352751
99608 1021541,97989,1021542
99627 14974,18803,2091
99642 135665,54834,54327
99647 130654,70122,124406
99648 1021562,1021563,1021561
9965 23659,22226,18979
99650 570276,1177641,41905
99657 1653805,6818
9966 14698,32597,37917
9968 18686,1462,2157
99708 2130,10127,2505
9973 216,61259,17140
99738 1090076,1153526,1153527
99749 14902,58293,2712
99758 105838,76941,76512
9978 33397,61363,3061
99785 139096,235712,567579
9980 17769,3798,2880
99819 1021737,1021736,1021738
99846 11163,75346,45469
99859 4935,9045,41381
99863 33022,14565,95314
9987 24695,586,7676
99875 150603,544283,13897
99879 1187562,1187561,1187563
99880 54233,123989,226027
99883 231170,37584,222589
99898 77081,33716,13576
9990 44149,51965,61556
99904 59297,1022659,133046
99916 32312,231170,37583
9993 1331,112,4783
9994 39949,61674,1905
9997 61776,5693,61779
99977 6844,8231,67764
[ss14396@hlog-1 MoviesToCastList]$


anujdhawan — ss14396@hlog-1:~/realtimeProject/MoviesToCastList — ssh ss14396@peel.hpc.nyu.edu — 204×56
FILE: Number of write operations=0
HDFS: Number of bytes read=190028928
HDFS: Number of bytes written=712017
HDFS: Number of read operations=11
HDFS: Number of large read operations=0
HDFS: Number of write operations=2
HDFS: Number of bytes read erasure-coded=0
Job Counters
Launched map tasks=2
Launched reduce tasks=1
Rack-local map tasks=2
Total time spent by all maps in occupied slots (ms)=40672
Total time spent by all reduces in occupied slots (ms)=13008
Total time spent by all map tasks (ms)=10168
Total time spent by all reduce tasks (ms)=2168
Total vcore-milliseconds taken by all map tasks=10168
Total vcore-milliseconds taken by all reduce tasks=2168
Total megabyte-milliseconds taken by all map tasks=41648128
Total megabyte-milliseconds taken by all reduce tasks=13320192
Map-Reduce Framework
Map input records=45477
Map output records=84352
Map output bytes=1057629
Map output materialized bytes=782960
Input split bytes=258
Combine input records=0
Combine output records=0
Reduce input groups=29259
Reduce shuffle bytes=782960
Reduce input records=84352
Reduce output records=29259
Spilled Records=168704
Shuffled Maps =2
Failed Shuffles=0
Merged Map outputs=2
GC time elapsed (ms)=208
CPU time spent (ms)=13950
Physical memory (bytes) snapshot=2241622016
Virtual memory (bytes) snapshot=11167272960
Total committed heap usage (bytes)=3393716224
Peak Map Physical memory (bytes)=935444480
Peak Map Virtual memory (bytes)=3717963776
Peak Reduce Physical memory (bytes)=466284544
Peak Reduce Virtual memory (bytes)=3741392896
Shuffle Errors
BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0
File Input Format Counters
Bytes Read=190028670
File Output Format Counters
Bytes Written=712017
[ss14396@hlog-1 MoviesToCastList]$


anujdhawan — ss14396@hlog-1:~/realtimeProject/MoviesToCastList — ssh ss14396@peel.hpc.nyu.edu — 204×56
at org.apache.hadoop.util.RunJar.main(RunJar.java:227)
[ss14396@hlog-1 MoviesToCastList]$ hadoop jar getMoviesToCastList.jar GetMoviesToCastListDriver /user/ss14396/rproject/credits.tsv /user/ss14396/rproject/moviesToCastListOutput
WARNING: Use "yarn jar" to launch YARN applications.
22/04/21 13:55:07 INFO client.RMProxy: Connecting to ResourceManager at horton.hpc.nyu.edu/10.32.35.134:8032
22/04/21 13:55:07 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to r
22/04/21 13:55:07 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /user/ss14396/.staging/job_1648648882306_27736
22/04/21 13:55:08 INFO input.FileInputFormat: Total input files to process : 1
22/04/21 13:55:08 INFO mapreduce.JobSubmitter: number of splits:2
22/04/21 13:55:08 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
22/04/21 13:55:08 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1648648882306_27736
22/04/21 13:55:08 INFO mapreduce.JobSubmitter: Executing with tokens: []
22/04/21 13:55:08 INFO conf.Configuration: resource-types.xml not found
22/04/21 13:55:08 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
22/04/21 13:55:08 INFO impl.YarnClientImpl: Submitted application application_1648648882306_27736
22/04/21 13:55:08 INFO mapreduce.Job: The url to track the job: http://horton.hpc.nyu.edu:8088/proxy/application_1648648882306_27736/
22/04/21 13:55:08 INFO mapreduce.Job: Running job: job_1648648882306_27736
22/04/21 13:55:13 INFO mapreduce.Job: Job job_1648648882306_27736 running in uber mode : false
22/04/21 13:55:13 INFO mapreduce.Job: map 0% reduce 0%
22/04/21 13:55:19 INFO mapreduce.Job: map 50% reduce 0%
22/04/21 13:55:21 INFO mapreduce.Job: map 100% reduce 0%
22/04/21 13:55:26 INFO mapreduce.Job: map 100% reduce 100%
22/04/21 13:55:27 INFO mapreduce.Job: Job job_1648648882306_27736 completed successfully
22/04/21 13:55:27 INFO mapreduce.Job: Counters: 54
File System Counters
FILE: Number of bytes read=779995
FILE: Number of bytes written=2227034
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=190028928
HDFS: Number of bytes written=712017
HDFS: Number of read operations=11
HDFS: Number of large read operations=0
HDFS: Number of write operations=2
HDFS: Number of bytes read erasure-coded=0
Job Counters
Launched map tasks=2
Launched reduce tasks=1
Rack-local map tasks=2
Total time spent by all maps in occupied slots (ms)=40672
Total time spent by all reduces in occupied slots (ms)=13008
Total time spent by all map tasks (ms)=10168
Total time spent by all reduce tasks (ms)=2168
Total vcore-milliseconds taken by all map tasks=10168
Total vcore-milliseconds taken by all reduce tasks=2168
Total megabyte-milliseconds taken by all map tasks=41648128
Total megabyte-milliseconds taken by all reduce tasks=13320192
Map-Reduce Framework
Map input records=45477
Map output records=84352
Map output bytes=1057629
Map output materialized bytes=782960
Input split bytes=258
Combine input records=0
Combine output records=0
Reduce input groups=29259