Task: A4 Data Preparation
Name: Edgardo Andres Panza Penso
Date: 10/Jan/2020

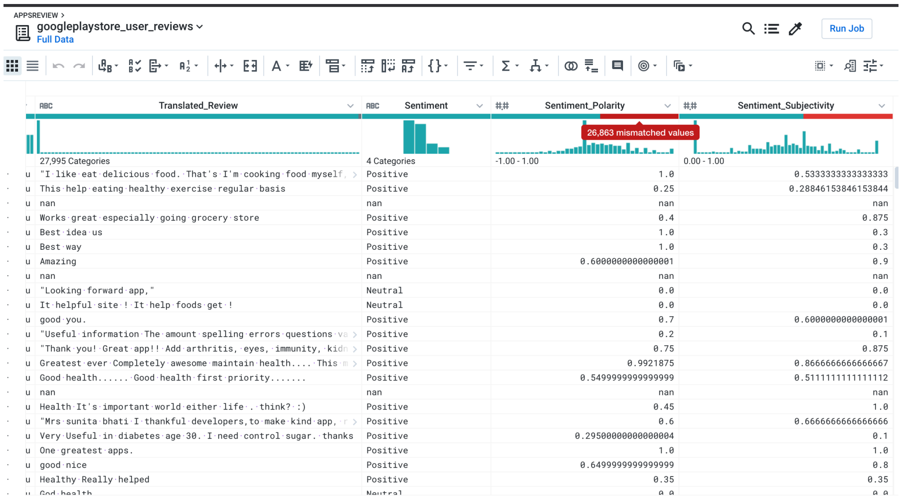
In the git repository you can find:

- Folder with datasets that i use for testing TRIFACTA
- LinksTask4.txt file with the links for the data cleaning training
- Readme.md
- A4_DataPreparation.pdf

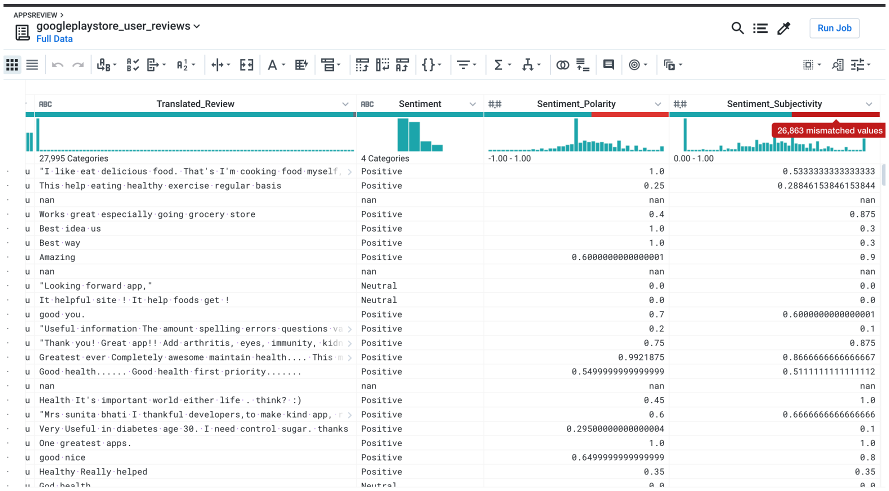
https://github.com/edpape007/DataPreparationExercise

TRIFACTA EXERCISE

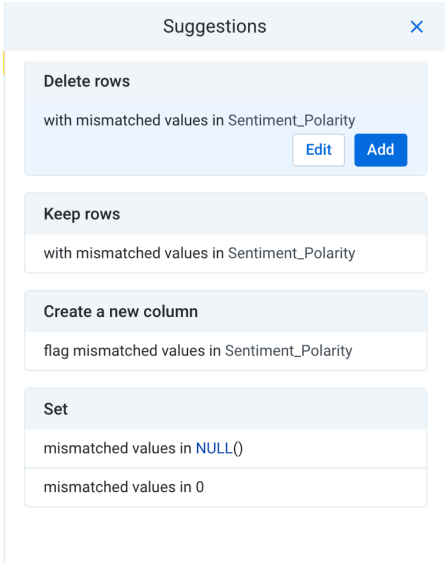
- First flow:



Translated_Review	Sentiment	Sentiment_Polarity	Sentiment_Subjectivity
27,995 Categories	4 Categories	-1.00 - 1.00	0.00 - 1.00
"I like eat delicious food. That's I'm cooking food myself,	Positive	1.0	0.5333333333333333
This help eating healthy exercise regular basis	Positive	0.25	0.28846153846153844
nan	nan	nan	nan
Works great especially going grocery store	Positive	0.4	0.875
Best idea us	Positive	1.0	0.3
Best way	Positive	1.0	0.3
Amazing	Positive	0.6000000000000001	0.9
nan	nan	nan	nan
"Looking forward app,"	Neutral	0.0	0.0
It helpful site ! It help foods get !	Neutral	0.0	0.0
good you.	Positive	0.7	0.6000000000000001
"Useful information The amount spelling errors questions va	Positive	0.2	0.1
"Thank you! Great app!! Add arthritis, eyes, immunity, kidn	Positive	0.75	0.875
Greatest ever Completely awesome maintain health.... This m	Positive	0.9921875	0.8666666666666667
Good health...... Good health first priority.......	Positive	0.5499999999999999	0.5111111111111112
nan	nan	nan	nan
Health It's important world either life . think? :)	Positive	0.45	1.0
"Mrs sunita bhati I thankful developers,to make kind app, r	Positive	0.6	0.6666666666666666
Very Useful in diabetes age 30. I need control sugar. thanks	Positive	0.29500000000000004	0.1
One greatest apps.	Positive	1.0	1.0
good nice	Positive	0.6499999999999999	0.8
Healthy Really helped	Positive	0.35	0.35
God health	Neutral	0.0	0.0



Translated_Review	Sentiment	Sentiment_Polarity	Sentiment_Subjectivity
27,995 Categories	4 Categories	-1.00 - 1.00	0.00 - 1.00
"I like eat delicious food. That's I'm cooking food myself,	Positive	1.0	0.5333333333333333
This help eating healthy exercise regular basis	Positive	0.25	0.28846153846153844
nan	nan	nan	nan
Works great especially going grocery store	Positive	0.4	0.875
Best idea us	Positive	1.0	0.3
Best way	Positive	1.0	0.3
Amazing	Positive	0.6000000000000001	0.9
nan	nan	nan	nan
"Looking forward app,"	Neutral	0.0	0.0
It helpful site ! It help foods get !	Neutral	0.0	0.0
good you.	Positive	0.7	0.6000000000000001
"Useful information The amount spelling errors questions va	Positive	0.2	0.1
"Thank you! Great app!! Add arthritis, eyes, immunity, kidn	Positive	0.75	0.875
Greatest ever Completely awesome maintain health.... This m	Positive	0.9921875	0.8666666666666667
Good health...... Good health first priority.......	Positive	0.5499999999999999	0.5111111111111112
nan	nan	nan	nan
Health It's important world either life . think? :)	Positive	0.45	1.0
"Mrs sunita bhati I thankful developers,to make kind app, r	Positive	0.6	0.6666666666666666
Very Useful in diabetes age 30. I need control sugar. thanks	Positive	0.29500000000000004	0.1
One greatest apps.	Positive	1.0	1.0
good nice	Positive	0.6499999999999999	0.8
Healthy Really helped	Positive	0.35	0.35
God health	Neutral	0.0	0.0



Suggestions ✕

Delete rows

with mismatched values in Sentiment_Polarity

Edit Add

Keep rows

with mismatched values in Sentiment_Polarity

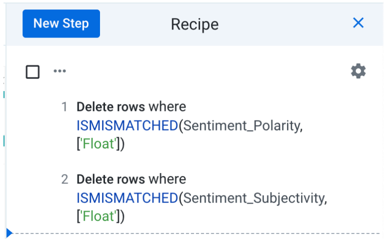
Create a new column

flag mismatched values in Sentiment_Polarity

Set

mismatched values in NULL()

mismatched values in 0

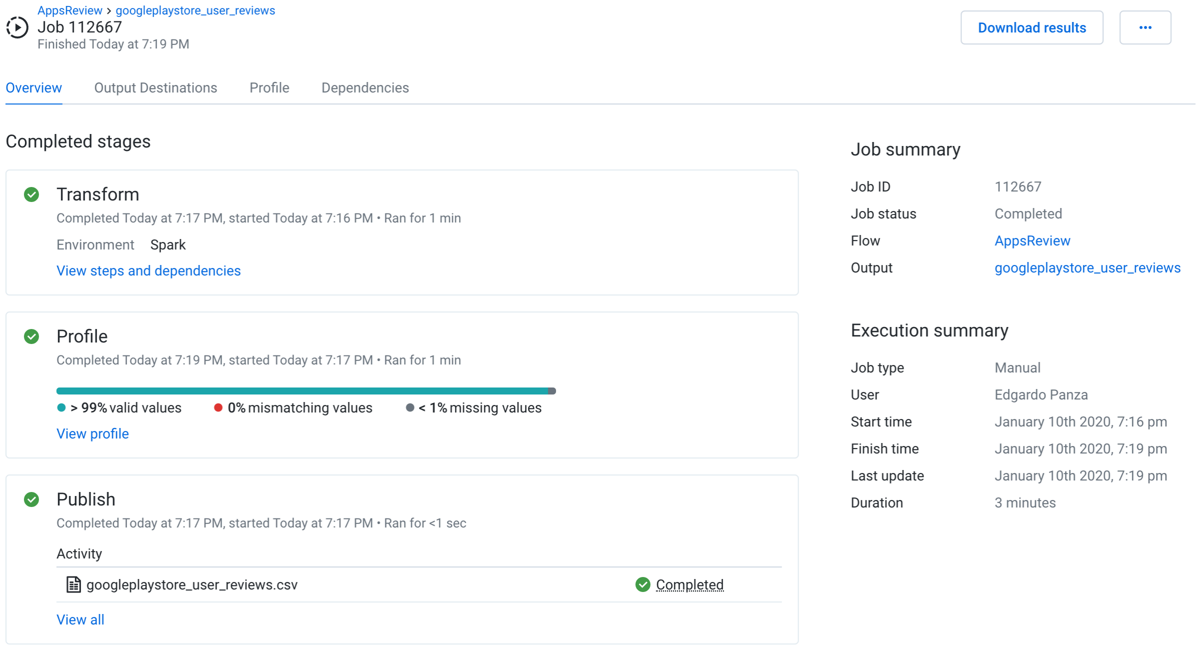


New Step Recipe ✕

□ ···

1 **Delete rows** where ISMISMATCHED(Sentiment_Polarity, ['Float'])

2 **Delete rows** where ISMISMATCHED(Sentiment_Subjectivity, ['Float'])



AppsReview > googleplaystore_user_reviews

Job 112667

Finished Today at 7:19 PM

Download results ···

Overview Output Destinations Profile Dependencies

Completed stages

✓ Transform

Completed Today at 7:17 PM, started Today at 7:16 PM • Ran for 1 min

Environment Spark

View steps and dependencies

✓ Profile

Completed Today at 7:19 PM, started Today at 7:17 PM • Ran for 1 min

● > 99% valid values ● 0% mismatching values ● < 1% missing values

View profile

✓ Publish

Completed Today at 7:17 PM, started Today at 7:17 PM • Ran for <1 sec

Activity

googleplaystore_user_reviews.csv ✓ Completed

View all

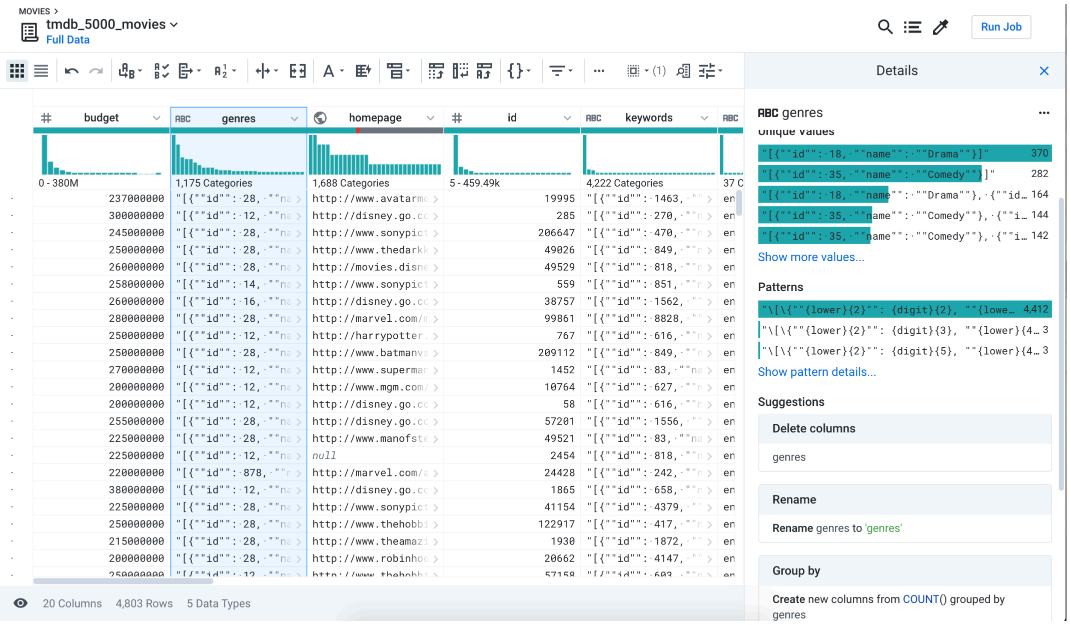
Job summary

Job ID	112667
Job status	Completed
Flow	AppsReview
Output	googleplaystore_user_reviews

Execution summary

Job type	Manual
User	Edgardo Panza
Start time	January 10th 2020, 7:16 pm
Finish time	January 10th 2020, 7:19 pm
Last update	January 10th 2020, 7:19 pm
Duration	3 minutes

- Second Flow
    - In this one I identify a pattern and create new column.
    - Delete the rows with missing values
    - Format the date column



MOVIES >
tmdb_5000_movies
Full Data

budget	genres	homepage	id	keywords	ABC
0 - 380M	1,175 Categories	1,688 Categories	5 - 459.49k	4,222 Categories	37 C
237000000	"[{""id"": 28, ""na	http://www.avatarmo	19995	"[{""id"": 1463, ""	en
300000000	"[{""id"": 12, ""na	http://disney.go.cc	285	"[{""id"": 270, ""n	en
245000000	"[{""id"": 28, ""na	http://www.sonypict	206647	"[{""id"": 470, ""n	en
250000000	"[{""id"": 28, ""na	http://www.thedarkk	49026	"[{""id"": 849, ""n	en
260000000	"[{""id"": 28, ""na	http://movies.disne	49529	"[{""id"": 818, ""n	en
258000000	"[{""id"": 14, ""na	http://www.sonypict	559	"[{""id"": 851, ""n	en
260000000	"[{""id"": 16, ""na	http://disney.go.cc	38757	"[{""id"": 1562, ""	en
280000000	"[{""id"": 28, ""na	http://marvel.com/m	99861	"[{""id"": 8828, ""	en
250000000	"[{""id"": 12, ""na	http://harrypotter.	767	"[{""id"": 616, ""n	en
250000000	"[{""id"": 28, ""na	http://www.batmanvs	209112	"[{""id"": 849, ""n	en
270000000	"[{""id"": 12, ""na	http://www.superman	1452	"[{""id"": 83, ""na	en
200000000	"[{""id"": 12, ""na	http://www.mgm.com/	10764	"[{""id"": 627, ""n	en
200000000	"[{""id"": 12, ""na	http://disney.go.cc	58	"[{""id"": 616, ""n	en
255000000	"[{""id"": 28, ""na	http://disney.go.cc	57201	"[{""id"": 1556, ""	en
225000000	"[{""id"": 28, ""na	http://www.manofste	49521	"[{""id"": 83, ""na	en
225000000	"[{""id"": 12, ""na	null	2454	"[{""id"": 818, ""n	en
220000000	"[{""id"": 878, ""n	http://marvel.com/a	24428	"[{""id"": 242, ""n	en
380000000	"[{""id"": 12, ""na	http://disney.go.cc	1865	"[{""id"": 658, ""n	en
225000000	"[{""id"": 28, ""na	http://www.sonypict	41154	"[{""id"": 4379, ""	en
250000000	"[{""id"": 28, ""na	http://www.thehobbi	122917	"[{""id"": 417, ""n	en
215000000	"[{""id"": 28, ""na	http://www.theamazi	1930	"[{""id"": 1872, ""	en
200000000	"[{""id"": 28, ""na	http://www.robinhoo	20662	"[{""id"": 4147, ""	en
250000000	"[{""id"": 12, ""na	http://www.thehobbi	57158	"[{""id"": 603, ""n	en

20 Columns 4,803 Rows 5 Data Types

Details

ABC genres

Unique values

"[{""id"": 18, ""name"": ""Drama""}]"	370
"[{""id"": 35, ""name"": ""Comedy""}]"	282
"[{""id"": 18, ""name"": ""Drama""}, {""id…	164
"[{""id"": 35, ""name"": ""Comedy""}, {""i…	144
"[{""id"": 35, ""name"": ""Comedy""}, {""i…	142

Show more values...

Patterns

"\[\{""{lower}{2}"": {digit}{2}, ""{lowe…	4,412
"\[\{""{lower}{2}"": {digit}{3}, ""{lower}{4…	3
"\[\{""{lower}{2}"": {digit}{5}, ""{lower}{4…	3

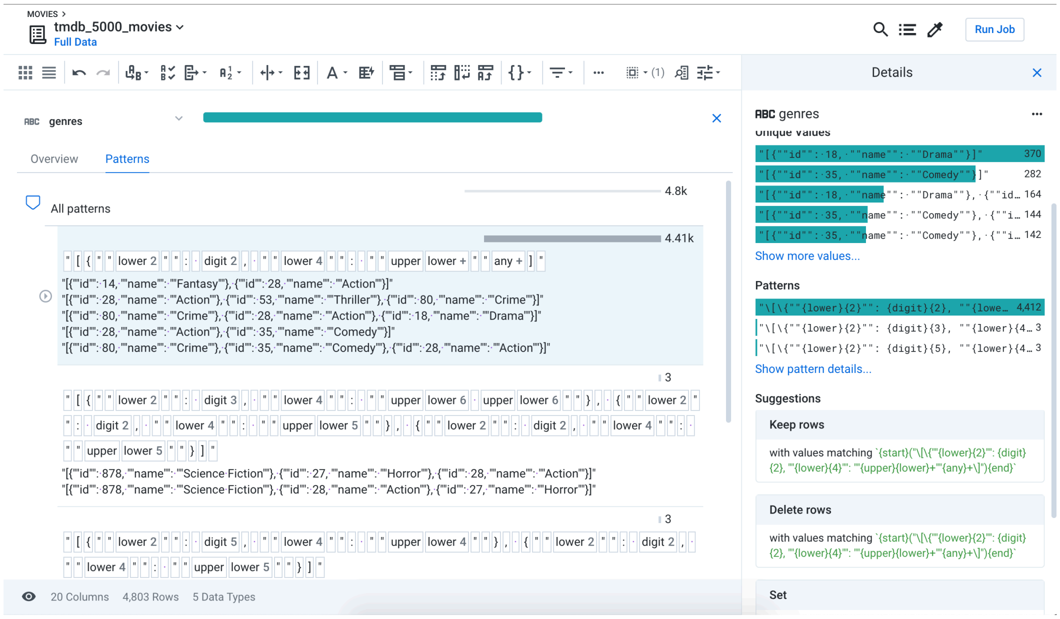
Show pattern details...

Suggestions

Delete columns
genres

Rename
Rename genres to 'genres'

Group by
Create new columns from COUNT() grouped by genres



MOVIES >
tmdb_5000_movies
Full Data

ABC genres

All patterns 4.8k

4.41k

" [{ " " lower 2 " " : digit 2 , " " lower 4 " " : " " upper lower + " " any +] "

"[{""id"": 14, ""name"": ""Fantasy""}, {""id"": 28, ""name"": ""Action""}]"
"[{""id"": 28, ""name"": ""Action""}, {""id"": 53, ""name"": ""Thriller""}, {""id"": 80, ""name"": ""Crime""}]"
"[{""id"": 80, ""name"": ""Crime""}, {""id"": 28, ""name"": ""Action""}, {""id"": 18, ""name"": ""Drama""}]"
"[{""id"": 28, ""name"": ""Action""}, {""id"": 35, ""name"": ""Comedy""}]"
"[{""id"": 80, ""name"": ""Crime""}, {""id"": 35, ""name"": ""Comedy""}, {""id"": 28, ""name"": ""Action""}]"

3

" [{ " " lower 2 " " : digit 3 , " " lower 4 " " : " " upper lower 6 upper lower 6 " " } , { " " lower 2 " " : digit 2 , " " lower 4 " " : " " upper lower 5 " " } , { " " lower 2 " " : digit 2 , " " lower 4 " " : " " upper lower 5 " " }] "

"[{""id"": 878, ""name"": ""Science Fiction""}, {""id"": 27, ""name"": ""Horror""}, {""id"": 28, ""name"": ""Action""}]"
"[{""id"": 878, ""name"": ""Science Fiction""}, {""id"": 28, ""name"": ""Action""}, {""id"": 27, ""name"": ""Horror""}]"

3

" [{ " " lower 2 " " : digit 5 , " " lower 4 " " : " " upper lower 4 " " } , { " " lower 2 " " : digit 2 , " " lower 4 " " : " " upper lower 5 " " }] "

20 Columns 4,803 Rows 5 Data Types

Details

ABC genres

Unique values

"[{""id"": 18, ""name"": ""Drama""}]"	370
"[{""id"": 35, ""name"": ""Comedy""}]"	282
"[{""id"": 18, ""name"": ""Drama""}, {""id…	164
"[{""id"": 35, ""name"": ""Comedy""}, {""i…	144
"[{""id"": 35, ""name"": ""Comedy""}, {""i…	142

Show more values...

Patterns

"\[\{""{lower}{2}"": {digit}{2}, ""{lowe…	4,412
"\[\{""{lower}{2}"": {digit}{3}, ""{lower}{4…	3
"\[\{""{lower}{2}"": {digit}{5}, ""{lower}{4…	3

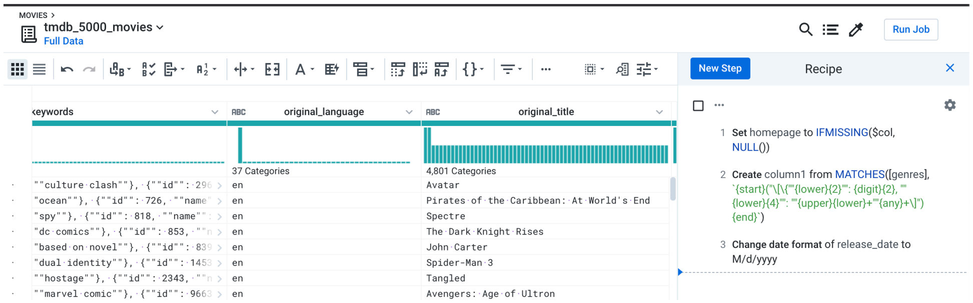
Show pattern details...

Suggestions

Keep rows
with values matching `{start}("\[\{""{lower}{2}"": {digit}{2}, ""{lower}{4}"": ""{upper}{lower}+""{any}+\]"){end}`

Delete rows
with values matching `{start}("\[\{""{lower}{2}"": {digit}{2}, ""{lower}{4}"": ""{upper}{lower}+""{any}+\]"){end}`

Set

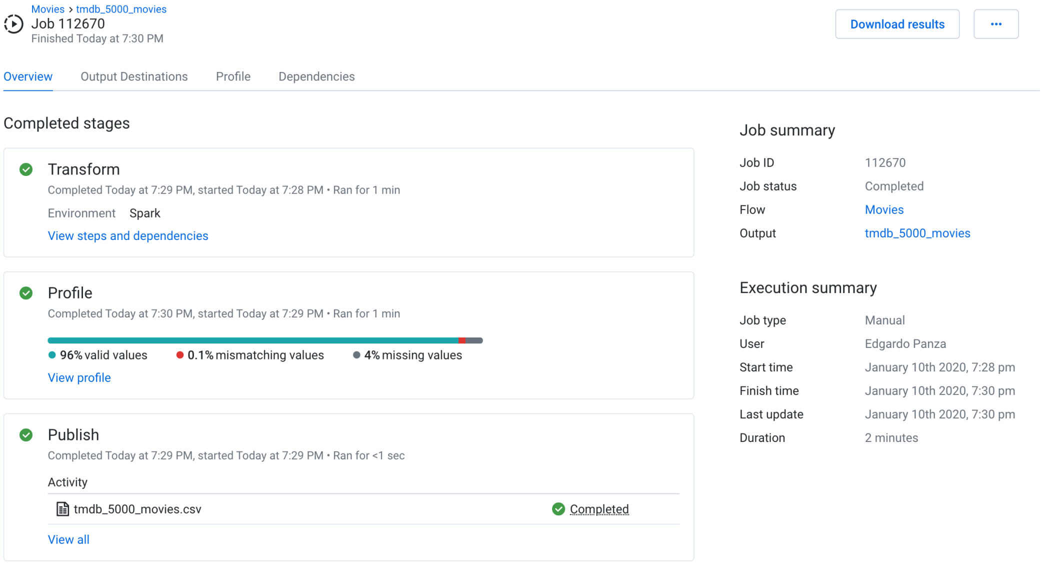


keywords	ABC original_language	ABC original_title
	37 Categories	4,801 Categories
""culture clash""}, {""id"": 296	en	Avatar
"ocean""}, {""id"": 726, ""name"	en	Pirates of the Caribbean: At World's End
"spy""}, {""id"": 818, ""name""	en	Spectre
"dc comics""}, {""id"": 853, ""n	en	The Dark Knight Rises
"based on novel""}, {""id"": 839	en	John Carter
"dual identity""}, {""id"": 1453	en	Spider-Man 3
""hostage""}, {""id"": 2343, ""n	en	Tangled
""marvel comic""}, {""id"": 9663	en	Avengers: Age of Ultron

1 **Set** homepage to IFMISSING($col, NULL())

2 **Create** column1 from MATCHES([genres], `{start}("\[\{""{lower}{2}"": {digit}{2}, ""{lower}{4}"": ""{upper}{lower}+""{any}+\]"){end}`)

3 **Change date format** of release_date to M/d/yyyy



Job 112670

Finished Today at 7:30 PM

Completed stages

Transform

Completed Today at 7:29 PM, started Today at 7:28 PM • Ran for 1 min

Environment Spark

View steps and dependencies

Profile

Completed Today at 7:30 PM, started Today at 7:29 PM • Ran for 1 min

● 96% valid values ● 0.1% mismatching values ● 4% missing values

View profile

Publish

Completed Today at 7:29 PM, started Today at 7:29 PM • Ran for <1 sec

Activity

tmdb_5000_movies.csv ✓ Completed

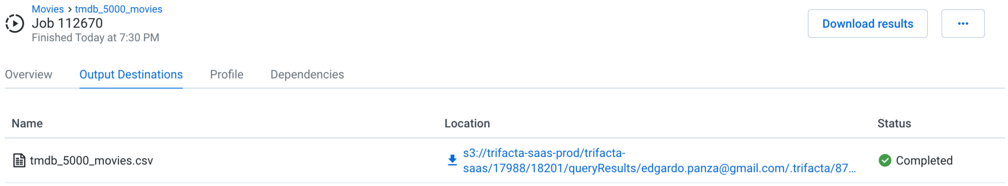
View all

Job summary

Job ID	112670
Job status	Completed
Flow	Movies
Output	tmdb_5000_movies

Execution summary

Job type	Manual
User	Edgardo Panza
Start time	January 10th 2020, 7:28 pm
Finish time	January 10th 2020, 7:30 pm
Last update	January 10th 2020, 7:30 pm
Duration	2 minutes



Job 112670

Finished Today at 7:30 PM

Name	Location	Status
tmdb_5000_movies.csv	s3://trifacta-saas-prod/trifacta-saas/17988/18201/queryResults/edgardo.panza@gmail.com/.trifacta/87…	✓ Completed