

download_WRDS_CIQ_transcripts_componenttext

May 27, 2025

1 Download Component Text Files from Capital IQ Transcripts

```
[1]: %matplotlib inline
# Do below if you want interactive matplotlib plot (). You can zoom in / zoom
    ↳out.
# %matplotlib notebook

# reloads modules automatically before entering the execution of code typed at
    ↳the IPython prompt
%load_ext autoreload
%autoreload 2
```

```
[2]: import wrds, os, re
import pandas as pd
from functions_gen import *
```

```
[3]: # setting up options
pd.set_option('display.memory_usage', 'deep')
pd.set_option('display.precision', 2)
pd.set_option('display.width', 240)
pd.set_option('display.max_rows', 4000)
pd.options.display.max_columns = None
pd.options.display.float_format = '{:,.4f}'.format
```

```
[17]: db = wrds.Connection(wrds_username = '#####')
```

Loading library list...
Done

```
[5]: db.list_tables(library='ciq_transcripts')
```

```
[5]: ['ciqtranscript',
      'ciqtranscriptcollectiontype',
      'ciqtranscriptcomponent',
      'ciqtranscriptcomponenttype',
      'ciqtranscriptdelayreason',
      'ciqtranscriptdelayreasontype',
      'ciqtranscriptperson',
```

```
'ciqtranscriptpresentationtype',
'ciqtranscriptspeakertype',
'wrds_transcript_detail',
'wrds_transcript_person']
```

```
[7]: wrds_transcript_detail = pd.read_pickle("../output/wrds_transcript_detail.pkl",
      ↪compression='zip')
```

```
[8]: wrds_transcript_detail.duplicated(subset=['transcriptid']).value_counts()
```

```
[8]: False    1543357
      True      32270
      Name: count, dtype: int64
```

```
[9]: transcripts = wrds_transcript_detail.drop_duplicates(subset=['transcriptid']).
      ↪sort_values(by='transcriptid')
```

```
[10]: transcripts['transcriptid'] = transcripts['transcriptid'].astype(int)
```

```
[11]: transcripts['quantile_30'], bins = pd.qcut(transcripts.transcriptid, q=30,
      ↪precision=0, retbins=True)
      transcripts['quantile_30'].value_counts(sort=False)
```

```
[11]: quantile_30
      (107.0, 73767.0]      51446
      (73767.0, 153556.0]    51445
      (153556.0, 273691.0]    51445
      (273691.0, 361010.0]    51445
      (361010.0, 454969.0]    51446
      (454969.0, 556695.0]    51445
      (556695.0, 660803.0]    51445
      (660803.0, 764934.0]    51445
      (764934.0, 866928.0]    51445
      (866928.0, 969849.0]    51446
      (969849.0, 1073219.0]   51445
      (1073219.0, 1196449.0]  51445
      (1196449.0, 1319492.0]  51445
      (1319492.0, 1416344.0]  51445
      (1416344.0, 1509145.0]  51446
      (1509145.0, 1602210.0]  51445
      (1602210.0, 1691225.0]  51445
      (1691225.0, 1783798.0]  51445
      (1783798.0, 1869604.0]  51445
      (1869604.0, 1949797.0]  51446
      (1949797.0, 2038333.0]  51445
      (2038333.0, 2127920.0]  51445
      (2127920.0, 2254900.0]  51445
```

```

(2254900.0, 2391959.0]    51445
(2391959.0, 2507111.0]    51446
(2507111.0, 2621232.0]    51445
(2621232.0, 2736944.0]    51445
(2736944.0, 2864493.0]    51445
(2864493.0, 3010496.0]    51445
(3010496.0, 3175090.0]    51446
Name: count, dtype: int64

```

```
[12]: bins.tolist()
```

```

[12]: [108.0,
       73767.2,
       153555.59999999998,
       273690.6,
       361010.19999999995,
       454969.0,
       556694.60000000001,
       660802.8,
       764933.79999999999,
       866928.39999999999,
       969849.0,
       1073219.2,
       1196448.6,
       1319491.6,
       1416343.8,
       1509145.0,
       1602210.2,
       1691224.6,
       1783797.6,
       1869603.7999999998,
       1949797.0,
       2038333.4,
       2127920.4,
       2254900.1999999997,
       2391958.80000000003,
       2507111.0,
       2621232.2,
       2736943.60000000006,
       2864493.2,
       3010495.8,
       3175090.0]

```

```

[13]: def get_component_sql(start, end, output, filenamesuf):
        df = db.raw_sql("select distinct * from ciq_transcripts.
        ↪ciqtranscriptcomponent where transcriptid>"+str(start)+" and
        ↪transcriptid<="+str(end))

```

```

    print(f"\n### Downloading transcript component, start transcriptid is {start}, end transcriptid is {end} ###\n")
    print("# descriptive statistics")
    print(df.info(), "\n")
    print(df.duplicated(subset=['transcriptid', 'transcriptcomponentid']).value_counts(), "\n")
    print(df.describe(), "\n")
    print("# saving zip compressed pickle as ciqtranscriptcomponent_"+filenamesuf)
    df.to_pickle(output+"/ciqtranscriptcomponent_"+filenamesuf+".pkl", compression="zip")

```

```

[15]: bins_test = [100, 500, 1000]
i=0
while i < len(bins.tolist())-1:
    print(f"\n##### PROCESSING BATCH {str(i+1).zfill(2)} #####\n")
    start = bins.tolist()[i]
    end = bins.tolist()[i+1]
    print(f"START TRANSCRIPTID is {start}, END TRANSCRIPTID is {end}")
    get_component_sql(start, end, "../output", str(i+1).zfill(2))
    i+=1

```

PROCESSING BATCH 01

START TRANSCRIPTID is 108.0, END TRANSCRIPTID is 73767.2

Downloading transcript component, start transcriptid is 108.0, end transcriptid is 73767.2

```

# descriptive statistics
<class 'pandas.core.frame.DataFrame'>
Index: 3860671 entries, 0 to 360670
Data columns (total 6 columns):
#   Column                                Dtype
---  -
0   transcriptcomponentid                 int64
1   transcriptid                         int64
2   componentorder                       int64
3   transcriptcomponenttypeid            int64
4   transcriptpersonid                   float64
5   componenttext                        object
dtypes: float64(1), int64(4), object(1)
memory usage: 3.0 GB
None

```

False 3860671

Name: count, dtype: int64

	transcriptcomponentid	transcriptid	componentorder
transcriptcomponenttypeid	transcriptpersonid		
count	3,860,671.0000	3,860,671.0000	3,860,671.0000
3,860,671.0000	3,858,041.0000		
mean	2,621,065.8149	32,601.0022	48.1643
3.9266	61,905.5083		
std	1,819,842.5235	21,497.7262	36.1338
1.4293	48,403.7134		
min	30,263.0000	109.0000	0.0000
1.0000	1.0000		
25%	1,111,051.5000	13,823.0000	20.0000
3.0000	10,150.0000		
50%	2,451,819.0000	29,456.0000	41.0000
4.0000	64,562.0000		
75%	3,985,939.5000	51,407.0000	68.0000
4.0000	103,027.0000		
max	15,377,492.0000	73,767.0000	373.0000
8.0000	196,301.0000		

saving zip compressed pickle as ciqtranscriptcomponent_01

PROCESSING BATCH 02

START TRANSCRIPTID is 73767.2, END TRANSCRIPTID is 153555.59999999998

Downloading transcript component, start transcriptid is 73767.2, end transcriptid is 153555.59999999998

descriptive statistics

<class 'pandas.core.frame.DataFrame'>

Index: 3377437 entries, 0 to 377436

Data columns (total 6 columns):

#	Column	Dtype
0	transcriptcomponentid	int64
1	transcriptid	int64
2	componentorder	int64
3	transcriptcomponenttypeid	int64
4	transcriptpersonid	float64
5	componenttext	object

dtypes: float64(1), int64(4), object(1)

memory usage: 2.6 GB

None

False 3377437

Name: count, dtype: int64

	transcriptcomponentid	transcriptid	componentorder
transcriptcomponenttypeid	transcriptpersonid		
count	3,377,437.0000	3,377,437.0000	3,377,437.0000
3,377,437.0000	3,376,702.0000		
mean	7,631,267.2001	113,240.0998	42.1185
3.8456	106,177.7108		
std	1,297,342.0954	23,201.1958	32.8631
1.3602	61,476.5819		
min	5,396,062.0000	73,768.0000	0.0000
1.0000	1.0000		
25%	6,533,580.0000	92,914.0000	17.0000
3.0000	92,196.0000		
50%	7,606,999.0000	112,921.0000	36.0000
4.0000	112,944.0000		
75%	8,750,299.0000	133,253.0000	60.0000
4.0000	154,165.0000		
max	15,685,916.0000	153,554.0000	309.0000
8.0000	229,782.0000		

saving zip compressed pickle as ciqtranscriptcomponent_02

PROCESSING BATCH 03

START TRANSCRIPTID is 153555.59999999998, END TRANSCRIPTID is 273690.6

Downloading transcript component, start transcriptid is 153555.59999999998, end transcriptid is 273690.6

descriptive statistics

<class 'pandas.core.frame.DataFrame'>

Index: 2964202 entries, 0 to 464201

Data columns (total 6 columns):

#	Column	Dtype
0	transcriptcomponentid	int64
1	transcriptid	int64
2	componentorder	int64
3	transcriptcomponenttypeid	int64
4	transcriptpersonid	float64
5	componenttext	object

dtypes: float64(1), int64(4), object(1)

memory usage: 2.5 GB

None

False 2964202

Name: count, dtype: int64

	transcriptcomponentid	transcriptid	componentorder
transcriptcomponenttypeid	transcriptpersonid		
count	2,964,202.0000	2,964,202.0000	2,964,202.0000
2,964,202.0000	2,963,944.0000		
mean	11,706,364.6702	201,975.4528	40.2813
3.7717	121,482.4105		
std	1,076,482.4195	35,552.5343	33.5562
1.3370	70,463.6364		
min	9,822,624.0000	153,558.0000	0.0000
1.0000	1.0000		
25%	10,781,162.2500	174,548.0000	15.0000
3.0000	93,407.0000		
50%	11,723,363.5000	193,993.0000	33.0000
4.0000	125,713.0000		
75%	12,627,832.7500	215,734.0000	57.0000
4.0000	175,886.0000		
max	17,307,751.0000	273,690.0000	365.0000
8.0000	232,149.0000		

saving zip compressed pickle as ciqtranscriptcomponent_03

PROCESSING BATCH 04

START TRANSCRIPTID is 273690.6, END TRANSCRIPTID is 361010.19999999995

Downloading transcript component, start transcriptid is 273690.6, end transcriptid is 361010.19999999995

descriptive statistics

<class 'pandas.core.frame.DataFrame'>

Index: 2935565 entries, 0 to 435564

Data columns (total 6 columns):

#	Column	Dtype
0	transcriptcomponentid	int64
1	transcriptid	int64
2	componentorder	int64
3	transcriptcomponenttypeid	int64
4	transcriptpersonid	int64
5	componenttext	object

dtypes: int64(5), object(1)

memory usage: 2.6 GB

None

False 2935565

Name: count, dtype: int64

transcriptcomponentid	transcriptid	componentorder
-----------------------	--------------	----------------

	transcriptcomponenttypeid	transcriptpersonid	
count	2,935,565.0000	2,935,565.0000	2,935,565.0000
2,935,565.0000	2,935,565.0000		
mean	15,412,861.4746	316,346.9278	39.4255
3.7713	133,783.0353		
std	1,071,888.8110	24,755.4354	33.2750
1.3371	76,684.6640		
min	13,576,724.0000	273,691.0000	0.0000
1.0000	1.0000		
25%	14,481,543.0000	294,039.0000	15.0000
3.0000	97,436.0000		
50%	15,400,004.0000	316,684.0000	32.0000
4.0000	140,476.0000		
75%	16,354,549.0000	337,577.0000	56.0000
4.0000	198,182.0000		
max	17,259,259.0000	361,007.0000	322.0000
8.0000	250,712.0000		

saving zip compressed pickle as ciqtranscriptcomponent_04

PROCESSING BATCH 05

START TRANSCRIPTID is 361010.19999999995, END TRANSCRIPTID is 454969.0

Downloading transcript component, start transcriptid is 361010.19999999995, end transcriptid is 454969.0

```
# descriptive statistics
<class 'pandas.core.frame.DataFrame'>
Index: 3082338 entries, 0 to 82337
Data columns (total 6 columns):
#   Column                                Dtype
---  -
0   transcriptcomponentid                 int64
1   transcriptid                         int64
2   componentorder                      int64
3   transcriptcomponenttypeid            int64
4   transcriptpersonid                  int64
5   componenttext                       object
dtypes: int64(5), object(1)
memory usage: 2.7 GB
None
```

```
False    3082338
Name: count, dtype: int64
```

```
transcriptcomponentid transcriptid componentorder
transcriptcomponenttypeid transcriptpersonid
```

count	3,082,338.0000	3,082,338.0000	3,082,338.0000
3,082,338.0000	3,082,338.0000		
mean	19,280,390.6647	407,861.7214	39.7122
3.7965	139,636.8032		
std	1,158,454.0201	27,606.1699	32.6318
1.3516	81,227.5767		
min	17,259,314.0000	361,011.0000	0.0000
1.0000	1.0000		
25%	18,263,644.2500	382,377.0000	15.0000
3.0000	97,937.0000		
50%	19,292,473.5000	408,511.0000	33.0000
4.0000	145,494.0000		
75%	20,255,800.7500	430,744.0000	56.0000
4.0000	210,152.0000		
max	21,295,783.0000	454,969.0000	361.0000
8.0000	265,754.0000		

saving zip compressed pickle as ciqtranscriptcomponent_05

PROCESSING BATCH 06

START TRANSCRIPTID is 454969.0, END TRANSCRIPTID is 556694.6000000001

Downloading transcript component, start transcriptid is 454969.0, end transcriptid is 556694.6000000001

descriptive statistics
<class 'pandas.core.frame.DataFrame'>
Index: 2927947 entries, 0 to 427946
Data columns (total 6 columns):

#	Column	Dtype
0	transcriptcomponentid	int64
1	transcriptid	int64
2	componentorder	int64
3	transcriptcomponenttypeid	int64
4	transcriptpersonid	int64
5	componenttext	object

dtypes: int64(5), object(1)
memory usage: 2.6 GB
None

False 2927947
Name: count, dtype: int64

	transcriptcomponentid	transcriptid	componentorder
	transcriptcomponenttypeid	transcriptpersonid	
count	2,927,947.0000	2,927,947.0000	2,927,947.0000

	2,927,947.0000	2,927,947.0000	
mean	23,322,145.4582	505,946.6489	38.2517
3.7781	151,068.2870		
std	1,224,918.3879	29,830.4367	32.8668
1.3608	87,730.5452		
min	21,293,732.0000	454,973.0000	0.0000
1.0000	1.0000		
25%	22,280,425.5000	482,466.0000	14.0000
3.0000	100,614.0000		
50%	23,223,773.0000	502,838.0000	31.0000
4.0000	159,897.0000		
75%	24,468,146.5000	534,250.0000	53.0000
4.0000	224,560.0000		
max	25,421,632.0000	556,694.0000	371.0000
8.0000	281,247.0000		

saving zip compressed pickle as ciqtranscriptcomponent_06

PROCESSING BATCH 07

START TRANSCRIPTID is 556694.6000000001, END TRANSCRIPTID is 660802.8

Downloading transcript component, start transcriptid is 556694.6000000001, end transcriptid is 660802.8

descriptive statistics

<class 'pandas.core.frame.DataFrame'>

Index: 2895989 entries, 0 to 395988

Data columns (total 6 columns):

#	Column	Dtype
0	transcriptcomponentid	int64
1	transcriptid	int64
2	componentorder	int64
3	transcriptcomponenttypeid	int64
4	transcriptpersonid	int64
5	componenttext	object

dtypes: int64(5), object(1)

memory usage: 2.5 GB

None

False 2895989

Name: count, dtype: int64

	transcriptcomponentid	transcriptid	componentorder
	transcriptcomponenttypeid	transcriptpersonid	
count	2,895,989.0000	2,895,989.0000	2,895,989.0000
2,895,989.0000	2,895,989.0000		

mean	27,425,199.9348	607,259.5390	38.3006
3.7267	159,480.1026		
std	1,110,842.3721	28,719.7696	33.5378
1.3379	91,706.7086		
min	25,421,633.0000	556,697.0000	0.0000
1.0000	1.0000		
25%	26,487,217.0000	582,872.0000	14.0000
3.0000	102,233.0000		
50%	27,419,805.0000	606,837.0000	31.0000
4.0000	169,525.0000		
75%	28,379,372.0000	630,374.0000	53.0000
4.0000	238,660.0000		
max	29,413,087.0000	660,802.0000	501.0000
8.0000	291,958.0000		

saving zip compressed pickle as ciqtranscriptcomponent_07

PROCESSING BATCH 08

START TRANSCRIPTID is 660802.8, END TRANSCRIPTID is 764933.7999999999

Downloading transcript component, start transcriptid is 660802.8, end transcriptid is 764933.7999999999

descriptive statistics

<class 'pandas.core.frame.DataFrame'>

Index: 2991468 entries, 0 to 491467

Data columns (total 6 columns):

#	Column	Dtype
0	transcriptcomponentid	int64
1	transcriptid	int64
2	componentorder	int64
3	transcriptcomponenttypeid	int64
4	transcriptpersonid	int64
5	componenttext	object

dtypes: int64(5), object(1)

memory usage: 2.4 GB

None

False 2991468

Name: count, dtype: int64

	transcriptcomponentid	transcriptid	componentorder
transcriptcomponenttypeid			transcriptpersonid
count	2,991,468.0000	2,991,468.0000	2,991,468.0000
2,991,468.0000	2,991,468.0000		
mean	31,447,119.5266	712,635.0691	39.2710

3.7733	167,886.5439		
std	1,195,735.1014	30,672.4828	34.9413
1.3508	96,838.8256		
min	29,403,041.0000	660,804.0000	0.0000
1.0000	1.0000		
25%	30,369,129.7500	684,991.0000	15.0000
3.0000	102,746.0000		
50%	31,492,922.5000	713,294.0000	31.0000
4.0000	179,392.0000		
75%	32,466,949.2500	738,829.0000	54.0000
4.0000	255,367.0000		
max	33,493,892.0000	764,932.0000	364.0000
8.0000	301,469.0000		

saving zip compressed pickle as ciqtranscriptcomponent_08

PROCESSING BATCH 09

START TRANSCRIPTID is 764933.7999999999, END TRANSCRIPTID is 866928.3999999999

Downloading transcript component, start transcriptid is 764933.7999999999, end transcriptid is 866928.3999999999

descriptive statistics

<class 'pandas.core.frame.DataFrame'>

Index: 2846694 entries, 0 to 346693

Data columns (total 6 columns):

#	Column	Dtype
0	transcriptcomponentid	int64
1	transcriptid	int64
2	componentorder	int64
3	transcriptcomponenttypeid	int64
4	transcriptpersonid	int64
5	componenttext	object

dtypes: int64(5), object(1)

memory usage: 2.3 GB

None

False 2846694

Name: count, dtype: int64

	transcriptcomponentid	transcriptid	componentorder
transcriptcomponenttypeid			
transcriptpersonid			
count	2,846,694.0000	2,846,694.0000	2,846,694.0000
	2,846,694.0000	2,846,694.0000	
mean	35,379,888.4145	815,299.3495	36.4142
3.7607	176,527.2438		

std	1,089,190.5925	29,421.4605	31.3164
1.3590	101,246.5536		
min	33,494,015.0000	764,935.0000	0.0000
1.0000	1.0000		
25%	34,442,937.2500	790,492.0000	14.0000
3.0000	104,145.0000		
50%	35,378,867.5000	814,831.0000	30.0000
4.0000	187,470.0000		
75%	36,329,506.7500	841,329.0000	51.0000
4.0000	270,156.0000		
max	37,269,245.0000	866,926.0000	350.0000
7.0000	310,007.0000		

saving zip compressed pickle as ciqtranscriptcomponent_09

PROCESSING BATCH 10

START TRANSCRIPTID is 866928.3999999999, END TRANSCRIPTID is 969849.0

Downloading transcript component, start transcriptid is 866928.3999999999, end transcriptid is 969849.0

descriptive statistics

<class 'pandas.core.frame.DataFrame'>

Index: 2907243 entries, 0 to 407242

Data columns (total 6 columns):

#	Column	Dtype
0	transcriptcomponentid	int64
1	transcriptid	int64
2	componentorder	int64
3	transcriptcomponenttypeid	int64
4	transcriptpersonid	int64
5	componenttext	object

dtypes: int64(5), object(1)

memory usage: 2.4 GB

None

False 2907243

Name: count, dtype: int64

	transcriptcomponentid	transcriptid	componentorder
transcriptcomponenttypeid			transcriptpersonid
count	2,907,243.0000	2,907,243.0000	2,907,243.0000
2,907,243.0000	2,907,243.0000		
mean	39,227,533.7693	919,079.9672	37.0616
3.7567	181,664.3649		
std	1,110,311.2393	29,456.0792	31.7072

1.3506	104,309.0524		
min	37,269,282.0000	866,929.0000	0.0000
1.0000	1.0000		
25%	38,275,845.5000	892,788.0000	14.0000
3.0000	104,853.0000		
50%	39,220,845.0000	920,365.0000	30.0000
4.0000	199,232.0000		
75%	40,182,097.5000	943,783.0000	52.0000
4.0000	279,153.0000		
max	41,159,666.0000	969,849.0000	448.0000
7.0000	318,097.0000		

saving zip compressed pickle as ciqtranscriptcomponent_10

PROCESSING BATCH 11

START TRANSCRIPTID is 969849.0, END TRANSCRIPTID is 1073219.2

Downloading transcript component, start transcriptid is 969849.0, end transcriptid is 1073219.2

descriptive statistics

<class 'pandas.core.frame.DataFrame'>

Index: 2781713 entries, 0 to 281712

Data columns (total 6 columns):

#	Column	Dtype
0	transcriptcomponentid	int64
1	transcriptid	int64
2	componentorder	int64
3	transcriptcomponenttypeid	int64
4	transcriptpersonid	int64
5	componenttext	object

dtypes: int64(5), object(1)

memory usage: 2.3 GB

None

False 2781713

Name: count, dtype: int64

	transcriptcomponentid	transcriptid	componentorder
transcriptcomponenttypeid			transcriptpersonid
count	2,781,713.0000	2,781,713.0000	2,781,713.0000
2,781,713.0000	2,781,713.0000		
mean	43,040,512.9413	1,022,009.6470	36.0509
3.7670	189,745.1170		
std	1,100,589.6625	30,039.6019	31.4006
1.3639	107,910.3661		

min	41,159,667.0000	969,850.0000	0.0000
1.0000	1.0000		
25%	42,063,713.0000	995,281.0000	14.0000
3.0000	106,344.0000		
50%	43,045,890.0000	1,021,843.0000	29.0000
4.0000	213,079.0000		
75%	43,944,232.0000	1,047,521.0000	50.0000
4.0000	290,289.0000		
max	44,959,623.0000	1,073,219.0000	342.0000
7.0000	325,282.0000		

saving zip compressed pickle as ciqtranscriptcomponent_11

PROCESSING BATCH 12

START TRANSCRIPTID is 1073219.2, END TRANSCRIPTID is 1196448.6

Downloading transcript component, start transcriptid is 1073219.2, end transcriptid is 1196448.6

descriptive statistics

<class 'pandas.core.frame.DataFrame'>

Index: 2773434 entries, 0 to 273433

Data columns (total 6 columns):

#	Column	Dtype
0	transcriptcomponentid	int64
1	transcriptid	int64
2	componentorder	int64
3	transcriptcomponenttypeid	int64
4	transcriptpersonid	int64
5	componenttext	object

dtypes: int64(5), object(1)

memory usage: 2.3 GB

None

False 2773434

Name: count, dtype: int64

	transcriptcomponentid	transcriptid	componentorder
transcriptcomponenttypeid		transcriptpersonid	
count	2,773,434.0000	2,773,434.0000	2,773,434.0000
2,773,434.0000	2,773,434.0000		
mean	47,067,045.2040	1,135,960.6966	35.5180
3.7660	195,305.7898		
std	1,236,889.7854	36,255.2608	30.6983
1.3657	110,902.5255		
min	44,959,624.0000	1,073,220.0000	0.0000

1.0000	1.0000		
25%	45,973,028.2500	1,104,240.0000	14.0000
3.0000	107,520.0000		
50%	47,096,126.5000	1,136,211.0000	29.0000
4.0000	219,991.0000		
75%	48,138,097.7500	1,167,464.0000	49.0000
4.0000	297,529.0000		
max	49,147,144.0000	1,196,447.0000	529.0000
7.0000	332,212.0000		

saving zip compressed pickle as ciqtranscriptcomponent_12

PROCESSING BATCH 13

START TRANSCRIPTID is 1196448.6, END TRANSCRIPTID is 1319491.6

Downloading transcript component, start transcriptid is 1196448.6, end transcriptid is 1319491.6

descriptive statistics

<class 'pandas.core.frame.DataFrame'>

Index: 2599783 entries, 0 to 99782

Data columns (total 6 columns):

#	Column	Dtype
0	transcriptcomponentid	int64
1	transcriptid	int64
2	componentorder	int64
3	transcriptcomponenttypeid	int64
4	transcriptpersonid	int64
5	componenttext	object

dtypes: int64(5), object(1)

memory usage: 2.2 GB

None

False 2599783

Name: count, dtype: int64

	transcriptcomponentid	transcriptid	componentorder
transcriptcomponenttypeid			
transcriptpersonid			
count	2,599,783.0000	2,599,783.0000	2,599,783.0000
2,599,783.0000	2,599,783.0000		
mean	51,095,480.0470	1,260,468.2371	35.9523
3.7110	207,655.4244		
std	1,119,109.2179	35,474.9760	34.4628
1.3438	116,416.5307		
min	49,147,145.0000	1,196,451.0000	0.0000
1.0000	1.0000		

25%	50,083,113.5000	1,229,183.0000	13.0000
3.0000	114,836.0000		
50%	51,150,961.0000	1,263,021.0000	28.0000
4.0000	238,023.0000		
75%	52,041,462.5000	1,288,304.0000	48.0000
4.0000	311,490.0000		
max	52,993,541.0000	1,319,491.0000	519.0000
7.0000	343,746.0000		

saving zip compressed pickle as ciqtranscriptcomponent_13

PROCESSING BATCH 14

START TRANSCRIPTID is 1319491.6, END TRANSCRIPTID is 1416343.8

Downloading transcript component, start transcriptid is 1319491.6, end transcriptid is 1416343.8

descriptive statistics

<class 'pandas.core.frame.DataFrame'>

Index: 2666138 entries, 0 to 166137

Data columns (total 6 columns):

#	Column	Dtype
0	transcriptcomponentid	int64
1	transcriptid	int64
2	componentorder	int64
3	transcriptcomponenttypeid	int64
4	transcriptpersonid	int64
5	componenttext	object

dtypes: int64(5), object(1)

memory usage: 2.2 GB

None

False 2666138

Name: count, dtype: int64

	transcriptcomponentid	transcriptid	componentorder
transcriptcomponenttypeid			
transcriptpersonid			
count	2,666,138.0000	2,666,138.0000	2,666,138.0000
2,666,138.0000	2,666,138.0000		
mean	54,698,242.7490	1,367,341.0493	35.6719
3.7406	213,097.8644		
std	1,006,059.7996	28,362.8442	32.0121
1.3562	119,866.8327		
min	52,993,542.0000	1,319,492.0000	0.0000
1.0000	1.0000		
25%	53,829,387.2500	1,342,549.0000	13.0000

3.0000	115,713.0000		
50%	54,684,611.5000	1,368,459.0000	28.0000
4.0000	251,966.0000		
75%	55,575,774.7500	1,392,530.0000	49.0000
4.0000	318,395.0000		
max	68,805,185.0000	1,416,343.0000	310.0000
7.0000	354,503.0000		

saving zip compressed pickle as ciqtranscriptcomponent_14

PROCESSING BATCH 15

START TRANSCRIPTID is 1416343.8, END TRANSCRIPTID is 1509145.0

Downloading transcript component, start transcriptid is 1416343.8, end transcriptid is 1509145.0

descriptive statistics

<class 'pandas.core.frame.DataFrame'>

Index: 2520655 entries, 0 to 20654

Data columns (total 6 columns):

#	Column	Dtype
0	transcriptcomponentid	int64
1	transcriptid	int64
2	componentorder	int64
3	transcriptcomponenttypeid	int64
4	transcriptpersonid	int64
5	componenttext	object

dtypes: int64(5), object(1)

memory usage: 2.1 GB

None

False 2520655

Name: count, dtype: int64

	transcriptcomponentid	transcriptid	componentorder
transcriptcomponenttypeid		transcriptpersonid	
count	2,520,655.0000	2,520,655.0000	2,520,655.0000
2,520,655.0000	2,520,655.0000		
mean	58,031,380.6769	1,461,196.6839	37.1994
3.6951	219,606.5878		
std	929,644.9191	26,486.2809	38.9196
1.3419	123,157.2204		
min	56,423,688.0000	1,416,344.0000	0.0000
1.0000	1.0000		
25%	57,239,805.5000	1,438,432.0000	12.0000
3.0000	117,099.0000		

50%	58,035,605.0000	1,460,403.0000	28.0000
4.0000	260,808.0000		
75%	58,807,136.5000	1,483,004.0000	49.0000
4.0000	327,416.0000		
max	59,692,646.0000	1,509,145.0000	630.0000
7.0000	363,978.0000		

saving zip compressed pickle as ciqtranscriptcomponent_15

PROCESSING BATCH 16

START TRANSCRIPTID is 1509145.0, END TRANSCRIPTID is 1602210.2

Downloading transcript component, start transcriptid is 1509145.0, end transcriptid is 1602210.2

descriptive statistics

<class 'pandas.core.frame.DataFrame'>

Index: 2655948 entries, 0 to 155947

Data columns (total 6 columns):

#	Column	Dtype
0	transcriptcomponentid	int64
1	transcriptid	int64
2	componentorder	int64
3	transcriptcomponenttypeid	int64
4	transcriptpersonid	int64
5	componenttext	object

dtypes: int64(5), object(1)

memory usage: 2.1 GB

None

False 2655948

Name: count, dtype: int64

	transcriptcomponentid	transcriptid	componentorder
transcriptcomponenttypeid			
transcriptpersonid			
count	2,655,948.0000	2,655,948.0000	2,655,948.0000
2,655,948.0000	2,655,948.0000		
mean	61,410,048.2943	1,555,210.3998	36.4215
3.7759	223,194.1829		
std	1,016,176.0157	26,893.1437	33.8144
1.3637	126,227.9742		
min	59,692,724.0000	1,509,148.0000	0.0000
1.0000	1.0000		
25%	60,548,381.7500	1,532,377.0000	13.0000
3.0000	119,717.0000		
50%	61,326,014.5000	1,553,977.0000	28.0000

```

4.0000      270,387.0000
75%         62,333,154.2500 1,579,858.0000      49.0000
4.0000      331,263.0000
max          63,182,685.0000 1,602,210.0000      418.0000
7.0000      372,987.0000

```

```
# saving zip compressed pickle as ciqtranscriptcomponent_16
```

```
##### PROCESSING BATCH 17 #####
```

```
START TRANSCRIPTID is 1602210.2, END TRANSCRIPTID is 1691224.6
```

```
### Downloading transcript component, start transcriptid is 1602210.2, end
transcriptid is 1691224.6 ###
```

```
# descriptive statistics
```

```
<class 'pandas.core.frame.DataFrame'>
```

```
Index: 2674398 entries, 0 to 174397
```

```
Data columns (total 6 columns):
```

```

#      Column                      Dtype
---  -
0      transcriptcomponentid      int64
1      transcriptid                int64
2      componentorder             int64
3      transcriptcomponenttypeid   int64
4      transcriptpersonid         int64
5      componenttext              object

```

```
dtypes: int64(5), object(1)
```

```
memory usage: 2.3 GB
```

```
None
```

```
False      2674398
```

```
Name: count, dtype: int64
```

```

      transcriptcomponentid  transcriptid  componentorder
transcriptcomponenttypeid  transcriptpersonid
count      2,674,398.0000 2,674,398.0000  2,674,398.0000
2,674,398.0000      2,674,398.0000
mean        64,866,761.0926 1,646,335.5350      37.7060
3.7131        235,652.4039
std          972,034.9661      24,848.9620      36.3760
1.3190        127,832.2937
min          63,182,686.0000 1,602,211.0000      0.0000
1.0000          1.0000
25%          64,024,968.2500 1,625,664.0000      13.0000
3.0000          139,436.0000
50%          64,866,062.5000 1,645,850.0000      29.0000
4.0000          287,952.0000

```

75%	65,706,471.7500	1,667,052.0000	50.0000
4.0000	340,282.0000		
max	66,546,381.0000	1,691,223.0000	401.0000
7.0000	383,240.0000		

saving zip compressed pickle as ciqtranscriptcomponent_17

PROCESSING BATCH 18

START TRANSCRIPTID is 1691224.6, END TRANSCRIPTID is 1783797.6

Downloading transcript component, start transcriptid is 1691224.6, end transcriptid is 1783797.6

descriptive statistics

<class 'pandas.core.frame.DataFrame'>

Index: 2635671 entries, 0 to 135670

Data columns (total 6 columns):

#	Column	Dtype
0	transcriptcomponentid	int64
1	transcriptid	int64
2	componentorder	int64
3	transcriptcomponenttypeid	int64
4	transcriptpersonid	int64
5	componenttext	object

dtypes: int64(5), object(1)

memory usage: 2.2 GB

None

False 2635671

Name: count, dtype: int64

	transcriptcomponentid	transcriptid	componentorder
transcriptcomponenttypeid			
transcriptpersonid			
count	2,635,671.0000	2,635,671.0000	2,635,671.0000
2,635,671.0000	2,635,671.0000		
mean	68,228,926.7571	1,740,167.0864	37.8543
3.7190	235,067.6876		
std	941,157.0579	26,686.5382	37.7053
1.3333	129,845.4914		
min	66,546,382.0000	1,691,227.0000	0.0000
1.0000	1.0000		
25%	67,422,673.5000	1,717,738.0000	13.0000
3.0000	132,528.0000		
50%	68,253,042.0000	1,742,125.0000	28.0000
4.0000	282,441.0000		
75%	69,028,725.5000	1,762,428.0000	50.0000

```

4.0000      342,606.0000
max         69,851,159.0000 1,783,797.0000      429.0000
7.0000      396,324.0000

```

```
# saving zip compressed pickle as ciqtranscriptcomponent_18
```

```
##### PROCESSING BATCH 19 #####
```

```
START TRANSCRIPTID is 1783797.6, END TRANSCRIPTID is 1869603.7999999998
```

```
### Downloading transcript component, start transcriptid is 1783797.6, end
transcriptid is 1869603.7999999998 ###
```

```
# descriptive statistics
```

```
<class 'pandas.core.frame.DataFrame'>
```

```
Index: 2635867 entries, 0 to 135866
```

```
Data columns (total 6 columns):
```

#	Column	Dtype
0	transcriptcomponentid	int64
1	transcriptid	int64
2	componentorder	int64
3	transcriptcomponenttypeid	int64
4	transcriptpersonid	int64
5	componenttext	object

```
dtypes: int64(5), object(1)
```

```
memory usage: 2.2 GB
```

```
None
```

```
False      2635867
```

```
Name: count, dtype: int64
```

	transcriptcomponentid	transcriptid	componentorder
count	2,635,867.0000	2,635,867.0000	2,635,867.0000
mean	71,496,341.1373	1,826,868.8797	36.4218
std	956,678.9711	25,041.8456	35.0709
min	69,851,160.0000	1,783,798.0000	0.0000
25%	70,671,680.5000	1,805,126.0000	13.0000
50%	71,467,202.0000	1,827,670.0000	28.0000
75%	72,334,419.5000	1,849,309.0000	49.0000
max	72,334,419.5000	1,849,309.0000	49.0000

```
max          73,154,399.0000 1,869,603.0000      429.0000
7.0000        408,038.0000
```

```
# saving zip compressed pickle as ciqtranscriptcomponent_19
```

```
##### PROCESSING BATCH 20 #####
```

```
START TRANSCRIPTID is 1869603.7999999998, END TRANSCRIPTID is 1949797.0
```

```
### Downloading transcript component, start transcriptid is 1869603.7999999998,
end transcriptid is 1949797.0 ###
```

```
# descriptive statistics
```

```
<class 'pandas.core.frame.DataFrame'>
```

```
Index: 2704208 entries, 0 to 204207
```

```
Data columns (total 6 columns):
```

#	Column	Dtype
0	transcriptcomponentid	int64
1	transcriptid	int64
2	componentorder	int64
3	transcriptcomponenttypeid	int64
4	transcriptpersonid	int64
5	componenttext	object

```
dtypes: int64(5), object(1)
```

```
memory usage: 2.3 GB
```

```
None
```

```
False      2704208
```

```
Name: count, dtype: int64
```

	transcriptcomponentid	transcriptid	componentorder
transcriptcomponenttypeid	transcriptpersonid		
count	2,704,208.0000	2,704,208.0000	2,704,208.0000
2,704,208.0000	2,704,208.0000		
mean	74,820,962.9910	1,906,985.4484	45.2285
3.7800	256,471.3101		
std	989,575.7530	22,444.3376	82.3934
1.3008	139,932.3725		
min	73,154,400.0000	1,869,604.0000	0.0000
1.0000	1.0000		
25%	73,941,719.7500	1,888,250.0000	13.0000
3.0000	158,823.0000		
50%	74,801,674.5000	1,904,864.0000	29.0000
4.0000	309,215.0000		
75%	75,690,071.2500	1,925,463.0000	53.0000
4.0000	367,773.0000		
max	76,519,254.0000	1,949,797.0000	1,867.0000

```

7.0000          427,747.0000

# saving zip compressed pickle as ciqtranscriptcomponent_20

##### PROCESSING BATCH 21 #####

START TRANSCRIPTID is 1949797.0, END TRANSCRIPTID is 2038333.4

### Downloading transcript component, start transcriptid is 1949797.0, end
transcriptid is 2038333.4 ###

# descriptive statistics
<class 'pandas.core.frame.DataFrame'>
Index: 2296307 entries, 0 to 296306
Data columns (total 6 columns):
 #   Column                                Dtype
---  -
 0   transcriptcomponentid                 int64
 1   transcriptid                         int64
 2   componentorder                      int64
 3   transcriptcomponenttypeid            int64
 4   transcriptpersonid                  int64
 5   componenttext                       object
dtypes: int64(5), object(1)
memory usage: 2.0 GB
None

False      2296307
Name: count, dtype: int64

      transcriptcomponentid  transcriptid  componentorder
transcriptcomponenttypeid  transcriptpersonid
count      2,296,307.0000  2,296,307.0000  2,296,307.0000
2,296,307.0000      2,296,307.0000
mean         78,111,246.8505  1,992,543.9815      34.7052
3.7410         260,719.4383
std          939,359.0829    27,068.2726      37.8166
1.3808         145,655.4956
min          76,519,281.0000  1,949,800.0000      0.0000
1.0000         1.0000
25%          77,308,240.5000  1,968,042.0000      11.0000
3.0000         150,050.0000
50%          78,131,895.0000  1,992,961.0000      25.0000
4.0000         311,152.0000
75%          78,885,477.5000  2,016,284.0000      46.0000
4.0000         376,656.0000
max          81,257,025.0000  2,038,333.0000      1,182.0000
7.0000         444,852.0000

```

```

# saving zip compressed pickle as ciqtranscriptcomponent_21

##### PROCESSING BATCH 22 #####

START TRANSCRIPTID is 2038333.4, END TRANSCRIPTID is 2127920.4

### Downloading transcript component, start transcriptid is 2038333.4, end
transcriptid is 2127920.4 ###

# descriptive statistics
<class 'pandas.core.frame.DataFrame'>
Index: 2614586 entries, 0 to 114585
Data columns (total 6 columns):
#   Column                                Dtype
---  -----
0   transcriptcomponentid                 int64
1   transcriptid                         int64
2   componentorder                      int64
3   transcriptcomponenttypeid            int64
4   transcriptpersonid                  int64
5   componenttext                       object
dtypes: int64(5), object(1)
memory usage: 2.2 GB
None

False      2614586
Name: count, dtype: int64

      transcriptcomponentid  transcriptid  componentorder
transcriptcomponenttypeid  transcriptpersonid
count      2,614,586.0000  2,614,586.0000   2,614,586.0000
2,614,586.0000      2,614,586.0000
mean        81,846,056.9347  2,084,311.4886           38.0891
3.8177        272,067.2373
std         1,140,826.9169    25,989.7566           37.6549
1.3474        148,950.7266
min         79,803,979.0000  2,038,335.0000           0.0000
1.0000           1.0000
25%         80,864,238.2500  2,061,609.0000           13.0000
3.0000        169,769.0000
50%         81,930,869.5000  2,086,973.5000           28.0000
4.0000        324,347.0000
75%         82,814,861.7500  2,106,481.0000           50.0000
4.0000        389,590.0000
max         83,769,326.0000  2,127,920.0000          402.0000
7.0000        458,769.0000

```

```

# saving zip compressed pickle as ciqtranscriptcomponent_22

##### PROCESSING BATCH 23 #####

START TRANSCRIPTID is 2127920.4, END TRANSCRIPTID is 2254900.1999999997

### Downloading transcript component, start transcriptid is 2127920.4, end
transcriptid is 2254900.1999999997 ###

# descriptive statistics
<class 'pandas.core.frame.DataFrame'>
Index: 2461070 entries, 0 to 461069
Data columns (total 6 columns):
#   Column                                Dtype
---  -----
0   transcriptcomponentid                 int64
1   transcriptid                         int64
2   componentorder                       int64
3   transcriptcomponenttypeid            int64
4   transcriptpersonid                   int64
5   componenttext                        object
dtypes: int64(5), object(1)
memory usage: 2.2 GB
None

False      2461070
Name: count, dtype: int64

      transcriptcomponentid  transcriptid  componentorder
transcriptcomponenttypeid  transcriptpersonid
count      2,461,070.0000  2,461,070.0000   2,461,070.0000
2,461,070.0000      2,461,070.0000
mean        85,611,750.4787  2,190,607.4963          34.4885
3.7949        277,783.3323
std         1,079,459.0696   40,236.0061          33.0596
1.3305        151,947.6406
min         83,769,327.0000  2,127,921.0000           0.0000
1.0000          1.0000
25%         84,692,699.2500  2,155,260.0000          12.0000
3.0000        170,721.0000
50%         85,589,528.5000  2,187,382.0000          26.0000
4.0000        326,078.0000
75%         86,543,960.7500  2,228,456.0000          46.0000
4.0000        396,938.0000
max         87,488,287.0000  2,254,899.0000         336.0000
7.0000        481,708.0000

# saving zip compressed pickle as ciqtranscriptcomponent_23

```

PROCESSING BATCH 24

START TRANSCRIPTID is 2254900.1999999997, END TRANSCRIPTID is 2391958.8000000003

Downloading transcript component, start transcriptid is 2254900.1999999997,
end transcriptid is 2391958.8000000003 ###

descriptive statistics

<class 'pandas.core.frame.DataFrame'>

Index: 2313245 entries, 0 to 313244

Data columns (total 6 columns):

#	Column	Dtype
0	transcriptcomponentid	int64
1	transcriptid	int64
2	componentorder	int64
3	transcriptcomponenttypeid	int64
4	transcriptpersonid	int64
5	componenttext	object

dtypes: int64(5), object(1)

memory usage: 2.0 GB

None

False 2313245

Name: count, dtype: int64

	transcriptcomponentid	transcriptid	componentorder
transcriptcomponenttypeid			
transcriptpersonid			
count	2,313,245.0000	2,313,245.0000	2,313,245.0000
2,313,245.0000	2,313,245.0000		
mean	89,350,337.7706	2,324,452.3962	34.1331
3.7821	286,283.0678		
std	1,064,839.0245	41,884.0998	34.3406
1.3747	162,054.2037		
min	87,488,307.0000	2,254,901.0000	0.0000
1.0000	1.0000		
25%	88,389,649.0000	2,284,782.0000	11.0000
3.0000	172,662.0000		
50%	89,428,875.0000	2,324,671.0000	25.0000
4.0000	332,822.0000		
75%	90,267,308.0000	2,362,623.0000	45.0000
4.0000	414,124.0000		
max	91,132,697.0000	2,391,954.0000	476.0000
7.0000	502,958.0000		

saving zip compressed pickle as ciqtranscriptcomponent_24

PROCESSING BATCH 25

START TRANSCRIPTID is 2391958.8000000003, END TRANSCRIPTID is 2507111.0

Downloading transcript component, start transcriptid is 2391958.8000000003,
end transcriptid is 2507111.0 ###

descriptive statistics

<class 'pandas.core.frame.DataFrame'>

Index: 2641809 entries, 0 to 141808

Data columns (total 6 columns):

#	Column	Dtype
0	transcriptcomponentid	int64
1	transcriptid	int64
2	componentorder	int64
3	transcriptcomponenttypeid	int64
4	transcriptpersonid	int64
5	componenttext	object

dtypes: int64(5), object(1)

memory usage: 2.3 GB

None

False 2641809

Name: count, dtype: int64

	transcriptcomponentid	transcriptid	componentorder
transcriptcomponenttypeid	transcriptpersonid		
count	2,641,809.0000	2,641,809.0000	2,641,809.0000
2,641,809.0000	2,641,809.0000		
mean	93,126,657.6356	2,452,934.0875	37.6460
3.7869	287,954.2878		
std	1,154,075.6196	33,666.2021	42.8172
1.3183	171,810.1858		
min	91,132,698.0000	2,391,960.0000	0.0000
1.0000	1.0000		
25%	92,116,497.0000	2,422,329.0000	13.0000
3.0000	149,306.0000		
50%	93,124,250.0000	2,456,651.0000	28.0000
4.0000	329,081.0000		
75%	94,131,503.0000	2,482,314.0000	49.0000
4.0000	422,794.0000		
max	95,082,979.0000	2,507,111.0000	1,182.0000
7.0000	528,519.0000		

saving zip compressed pickle as ciqtranscriptcomponent_25

OSError Traceback (most recent call last)

Cell In[15], line 8

```
6 end = bins.tolist()[i+1]
7 print(f"START TRANSCRIPTID is {start}, END TRANSCRIPTID is {end}")
----> 8 get_component_sql(start, end, "../output", str(i+1).zfill(2))
9 i+=1
```

Cell In[13], line 9, in get_component_sql(start, end, output, filenamesuf)

```
7 print(df.describe(), "\n")
8 print("# saving zip compressed pickle as_
↳ ciqtranscriptcomponent_" + filenamesuf)
----> 9_
↳ df.to_pickle(output + "/" + ciqtranscriptcomponent_ + filenamesuf + ".pkl", compression="zip")
```

File c:

```
↳ \Users\flakej\AppData\Local\Programs\Python\Python311\Lib\site-packages\pandas\core\generi
↳ py:2955, in NDFrame.to_pickle(self, path, compression, protocol,
↳ storage_options)
2903 """
2904 Pickle (serialize) object to file.
2905
2906 (...)
2951 4      4      9
2952 """ # noqa: E501
2953 from pandas.io.pickle import to_pickle
-> 2955 to_pickle(
2956     self,
2957     path,
2958     compression=compression,
2959     protocol=protocol,
2960     storage_options=storage_options,
2961 )
```

File c:

```
↳ \Users\flakej\AppData\Local\Programs\Python\Python311\Lib\site-packages\pandas\io\pickle.
↳ py:95, in to_pickle(obj, filepath_or_buffer, compression, protocol,
↳ storage_options)
92 if protocol < 0:
93     protocol = pickle.HIGHEST_PROTOCOL
----> 95 with get_handle(
96     filepath_or_buffer,
97     "wb",
98     compression=compression,
99     is_text=False,
100     storage_options=storage_options,
101 ) as handles:
```

```

102
↳ # letting pickle write directly to the buffer is more memory-efficient
103     pickle.dump(obj, handles.handle, protocol=protocol)

File c:
↳ \Users\flakej\AppData\Local\Programs\Python\Python311\Lib\site-packages\pandas\io\common.
py:138, in IOHandles.__exit__(self, *args)
    137 def __exit__(self, *args: Any) -> None:
--> 138     self.close()

File c:
↳ \Users\flakej\AppData\Local\Programs\Python\Python311\Lib\site-packages\pandas\io\common.
py:130, in IOHandles.close(self)
    128     self.created_handles.remove(self.handle)
    129     for handle in self.created_handles:
--> 130         handle.close()
    131     self.created_handles = []
    132     self.is_wrapped = False

File c:
↳ \Users\flakej\AppData\Local\Programs\Python\Python311\Lib\site-packages\pandas\io\common.
py:947, in _BufferedWriter.close(self)
    945     # error: "_BufferedWriter" has no attribute "buffer"
    946     with self.buffer: # type: ignore[attr-defined]
--> 947         self.write_to_buffer()
    948 else:
    949     # error: "_BufferedWriter" has no attribute "buffer"
    950     self.buffer.close() # type: ignore[attr-defined]

File c:
↳ \Users\flakej\AppData\Local\Programs\Python\Python311\Lib\site-packages\pandas\io\common.
py:1042, in _BytesZipFile.write_to_buffer(self)
    1039 def write_to_buffer(self) -> None:
    1040     # ZipFile needs a non-empty string
    1041     archive_name = self.archive_name or self.infer_filename() or "zip"
-> 1042     self.buffer.writestr(archive_name, self.getvalue())

File c:\Users\flakej\AppData\Local\Programs\Python\Python311\Lib\zipfile.py:
↳ 1835, in ZipFile.writestr(self, zinfo_or_arcname, data, compress_type,
↳ compresslevel)
    1833 with self._lock:
    1834     with self.open(zinfo, mode='w') as dest:
-> 1835         dest.write(data)

File c:\Users\flakej\AppData\Local\Programs\Python\Python311\Lib\zipfile.py:
↳ 1169, in _ZipWriteFile.write(self, data)
    1167     data = self._compressor.compress(data)
    1168     self._compress_size += len(data)
-> 1169     self._fileobj.write(data)

```

```
1170 return nbytes
```

```
OSError: [Errno 28] No space left on device
```

```
[19]: bins_test = [100, 500, 1000]
i=24
while i < len(bins.tolist())-1:
    print(f"\n##### PROCESSING BATCH {str(i+1).zfill(2)} #####\n")
    start = bins.tolist()[i]
    end = bins.tolist()[i+1]
    print(f"START TRANSCRIPTID is {start}, END TRANSCRIPTID is {end}")
    get_component_sql(start, end, "../output", str(i+1).zfill(2))
    i+=1
```

```
##### PROCESSING BATCH 25 #####
```

```
START TRANSCRIPTID is 2391958.8000000003, END TRANSCRIPTID is 2507111.0
```

```
### Downloading transcript component, start transcriptid is 2391958.8000000003,
end transcriptid is 2507111.0 ###
```

```
# descriptive statistics
<class 'pandas.core.frame.DataFrame'>
Index: 2641809 entries, 0 to 141808
Data columns (total 6 columns):
#   Column                                Dtype
---  -
0   transcriptcomponentid                 int64
1   transcriptid                         int64
2   componentorder                      int64
3   transcriptcomponenttypeid            int64
4   transcriptpersonid                  int64
5   componenttext                       object
dtypes: int64(5), object(1)
memory usage: 2.3 GB
None
```

```
False      2641809
Name: count, dtype: int64
```

```
transcriptcomponentid  transcriptid  componentorder
transcriptcomponenttypeid  transcriptpersonid
count      2,641,809.0000  2,641,809.0000  2,641,809.0000
2,641,809.0000      2,641,809.0000
mean      93,126,657.6356  2,452,934.0875      37.6460
3.7869      287,954.2878
```

std	1,154,075.6196	33,666.2021	42.8172
1.3183	171,810.1858		
min	91,132,698.0000	2,391,960.0000	0.0000
1.0000	1.0000		
25%	92,116,497.0000	2,422,329.0000	13.0000
3.0000	149,306.0000		
50%	93,124,250.0000	2,456,651.0000	28.0000
4.0000	329,081.0000		
75%	94,131,503.0000	2,482,314.0000	49.0000
4.0000	422,794.0000		
max	95,082,979.0000	2,507,111.0000	1,182.0000
7.0000	528,519.0000		

saving zip compressed pickle as ciqtranscriptcomponent_25

PROCESSING BATCH 26

START TRANSCRIPTID is 2507111.0, END TRANSCRIPTID is 2621232.2

Downloading transcript component, start transcriptid is 2507111.0, end transcriptid is 2621232.2

descriptive statistics

<class 'pandas.core.frame.DataFrame'>

Index: 2429020 entries, 0 to 429019

Data columns (total 6 columns):

#	Column	Dtype
0	transcriptcomponentid	int64
1	transcriptid	int64
2	componentorder	int64
3	transcriptcomponenttypeid	int64
4	transcriptpersonid	int64
5	componenttext	object

dtypes: int64(5), object(1)

memory usage: 2.2 GB

None

False 2429020

Name: count, dtype: int64

	transcriptcomponentid	transcriptid	componentorder
transcriptcomponenttypeid			transcriptpersonid
count	2,429,020.0000	2,429,020.0000	2,429,020.0000
	2,429,020.0000	2,429,020.0000	
mean	96,922,351.5274	2,563,875.0780	33.7410
3.7777	302,198.7041		
std	1,085,799.5088	32,700.3364	34.0591

```

1.3407      174,350.3846
min          95,082,980.0000 2,507,112.0000      0.0000
1.0000              1.0000
25%          95,966,529.7500 2,535,981.0000      12.0000
3.0000          180,743.0000
50%          96,901,624.5000 2,562,863.0000      25.0000
4.0000          339,012.0000
75%          97,853,135.2500 2,590,835.0000      44.0000
4.0000          443,341.0000
max          98,813,331.0000 2,621,232.0000      576.0000
7.0000          548,921.0000

```

saving zip compressed pickle as ciqtranscriptcomponent_26

PROCESSING BATCH 27

START TRANSCRIPTID is 2621232.2, END TRANSCRIPTID is 2736943.6000000006

Downloading transcript component, start transcriptid is 2621232.2, end transcriptid is 2736943.6000000006

descriptive statistics

<class 'pandas.core.frame.DataFrame'>

Index: 2482164 entries, 0 to 482163

Data columns (total 6 columns):

#	Column	Dtype
0	transcriptcomponentid	int64
1	transcriptid	int64
2	componentorder	int64
3	transcriptcomponenttypeid	int64
4	transcriptpersonid	int64
5	componenttext	object

dtypes: int64(5), object(1)

memory usage: 2.1 GB

None

False 2482164

Name: count, dtype: int64

	transcriptcomponentid	transcriptid	componentorder
count	2,482,164.0000	2,482,164.0000	2,482,164.0000
mean	100,712,925.4029	2,678,151.3138	34.9836
std	1,080,370.9994	33,460.0763	34.0807
min	95,082,980.0000	2,507,112.0000	0.0000
max	98,813,331.0000	2,621,232.0000	576.0000

```

min          98,813,332.0000 2,621,233.0000          0.0000
1.0000          1.0000
25%          99,784,546.7500 2,650,021.0000          12.0000
3.0000          182,536.0000
50%          100,707,509.5000 2,676,696.0000          26.0000
4.0000          357,443.0000
75%          101,638,368.2500 2,705,115.0000          46.0000
4.0000          475,861.0000
max          102,590,129.0000 2,736,942.0000          487.0000
7.0000          566,407.0000

```

```
# saving zip compressed pickle as ciqtranscriptcomponent_27
```

```
##### PROCESSING BATCH 28 #####
```

```
START TRANSCRIPTID is 2736943.6000000006, END TRANSCRIPTID is 2864493.2
```

```
### Downloading transcript component, start transcriptid is 2736943.6000000006,
end transcriptid is 2864493.2 ###
```

```
# descriptive statistics
```

```
<class 'pandas.core.frame.DataFrame'>
```

```
Index: 2346041 entries, 0 to 346040
```

```
Data columns (total 6 columns):
```

#	Column	Dtype
0	transcriptcomponentid	int64
1	transcriptid	int64
2	componentorder	int64
3	transcriptcomponenttypeid	int64
4	transcriptpersonid	int64
5	componenttext	object

```
dtypes: int64(5), object(1)
```

```
memory usage: 2.1 GB
```

```
None
```

```
False    2346041
```

```
Name: count, dtype: int64
```

	transcriptcomponentid	transcriptid	componentorder
transcriptcomponenttypeid	transcriptpersonid		
count	2,346,041.0000	2,346,041.0000	2,346,041.0000
2,346,041.0000	2,346,041.0000		
mean	104,406,474.4704	2,799,218.4229	33.2137
3.7805	324,308.6019		
std	1,096,281.0914	37,678.7882	35.0740
1.3444	189,541.7147		
min	102,590,354.0000	2,736,946.0000	0.0000

```

1.0000          1.0000
25%          103,438,591.0000  2,764,778.0000          11.0000
3.0000          182,536.0000
50%          104,340,195.0000  2,796,238.0000          25.0000
4.0000          357,395.0000
75%          105,394,725.0000  2,831,357.0000          44.0000
4.0000          489,673.0000
max          106,332,203.0000  2,864,492.0000          785.0000
7.0000          587,862.0000

```

```
# saving zip compressed pickle as ciqtranscriptcomponent_28
```

```
##### PROCESSING BATCH 29 #####
```

```
START TRANSCRIPTID is 2864493.2, END TRANSCRIPTID is 3010495.8
```

```
### Downloading transcript component, start transcriptid is 2864493.2, end
transcriptid is 3010495.8 ###
```

```
# descriptive statistics
```

```
<class 'pandas.core.frame.DataFrame'>
```

```
Index: 2445223 entries, 0 to 445222
```

```
Data columns (total 6 columns):
```

#	Column	Dtype
0	transcriptcomponentid	int64
1	transcriptid	int64
2	componentorder	int64
3	transcriptcomponenttypeid	int64
4	transcriptpersonid	int64
5	componenttext	object

```
dtypes: int64(5), object(1)
```

```
memory usage: 2.1 GB
```

```
None
```

```
False      2445223
```

```
Name: count, dtype: int64
```

	transcriptcomponentid	transcriptid	componentorder
transcriptcomponenttypeid			
transcriptpersonid			
count	2,445,223.0000	2,445,223.0000	2,445,223.0000
2,445,223.0000	2,445,223.0000		
mean	108,212,358.9812	2,938,149.8047	34.7975
3.8072	336,646.0368		
std	1,078,693.6367	42,688.7504	33.9411
1.3427	198,650.2733		
min	106,332,253.0000	2,864,494.0000	0.0000
1.0000	1.0000		

25%	107,276,250.5000	2,904,283.0000	12.0000
3.0000	182,536.0000		
50%	108,219,642.0000	2,935,553.0000	26.0000
4.0000	368,286.0000		
75%	109,171,545.5000	2,975,795.0000	46.0000
4.0000	509,554.0000		
max	110,027,993.0000	3,010,495.0000	534.0000
7.0000	607,951.0000		

saving zip compressed pickle as ciqtranscriptcomponent_29

PROCESSING BATCH 30

START TRANSCRIPTID is 3010495.8, END TRANSCRIPTID is 3175090.0

Downloading transcript component, start transcriptid is 3010495.8, end transcriptid is 3175090.0

descriptive statistics

<class 'pandas.core.frame.DataFrame'>

Index: 2432445 entries, 0 to 432444

Data columns (total 6 columns):

#	Column	Dtype
0	transcriptcomponentid	int64
1	transcriptid	int64
2	componentorder	int64
3	transcriptcomponenttypeid	int64
4	transcriptpersonid	int64
5	componenttext	object

dtypes: int64(5), object(1)

memory usage: 2.1 GB

None

False 2432445

Name: count, dtype: int64

	transcriptcomponentid	transcriptid	componentorder
count	2,432,445.0000	2,432,445.0000	2,432,445.0000
mean	111,802,782.4207	3,085,880.2633	34.5630
std	1,024,009.4746	47,285.7342	32.8821
min	110,027,994.0000	3,010,496.0000	0.0000
25%	110,923,672.0000	3,043,323.0000	12.0000

3.0000	182,795.0000		
50%	111,781,973.0000	3,085,717.0000	26.0000
4.0000	370,750.0000		
75%	112,681,302.0000	3,126,076.0000	46.0000
4.0000	522,723.0000		
max	113,613,786.0000	3,175,090.0000	411.0000
7.0000	623,418.0000		

saving zip compressed pickle as ciqtranscriptcomponent_30

```
[20]: del transcripts
```

Check that wrds_transcript_person files have data for all transcripts

```
[21]: import glob
from functools import reduce
filelist = glob.glob(os.path.abspath("../output/")+"\\ciqtranscriptcomponent_*")
# filelist.append(filelist.pop(0))
filelist
```

```
[21]: ['c:\\Users\\flakej\\Dropbox\\Research\\Data\\CIQ_Transcripts\\2024\\output\\ciq
transcriptcomponent_01.pkl',
'c:\\Users\\flakej\\Dropbox\\Research\\Data\\CIQ_Transcripts\\2024\\output\\ciq
transcriptcomponent_02.pkl',
'c:\\Users\\flakej\\Dropbox\\Research\\Data\\CIQ_Transcripts\\2024\\output\\ciq
transcriptcomponent_03.pkl',
'c:\\Users\\flakej\\Dropbox\\Research\\Data\\CIQ_Transcripts\\2024\\output\\ciq
transcriptcomponent_04.pkl',
'c:\\Users\\flakej\\Dropbox\\Research\\Data\\CIQ_Transcripts\\2024\\output\\ciq
transcriptcomponent_05.pkl',
'c:\\Users\\flakej\\Dropbox\\Research\\Data\\CIQ_Transcripts\\2024\\output\\ciq
transcriptcomponent_06.pkl',
'c:\\Users\\flakej\\Dropbox\\Research\\Data\\CIQ_Transcripts\\2024\\output\\ciq
transcriptcomponent_07.pkl',
'c:\\Users\\flakej\\Dropbox\\Research\\Data\\CIQ_Transcripts\\2024\\output\\ciq
transcriptcomponent_08.pkl',
'c:\\Users\\flakej\\Dropbox\\Research\\Data\\CIQ_Transcripts\\2024\\output\\ciq
transcriptcomponent_09.pkl',
'c:\\Users\\flakej\\Dropbox\\Research\\Data\\CIQ_Transcripts\\2024\\output\\ciq
transcriptcomponent_10.pkl',
'c:\\Users\\flakej\\Dropbox\\Research\\Data\\CIQ_Transcripts\\2024\\output\\ciq
transcriptcomponent_11.pkl',
'c:\\Users\\flakej\\Dropbox\\Research\\Data\\CIQ_Transcripts\\2024\\output\\ciq
transcriptcomponent_12.pkl',
'c:\\Users\\flakej\\Dropbox\\Research\\Data\\CIQ_Transcripts\\2024\\output\\ciq
transcriptcomponent_13.pkl',
'c:\\Users\\flakej\\Dropbox\\Research\\Data\\CIQ_Transcripts\\2024\\output\\ciq
transcriptcomponent_14.pkl',
```

```

'c:\\Users\\flakej\\Dropbox\\Research\\Data\\CIQ_Transcripts\\2024\\output\\ciq
transcriptcomponent_15.pkl',
'c:\\Users\\flakej\\Dropbox\\Research\\Data\\CIQ_Transcripts\\2024\\output\\ciq
transcriptcomponent_16.pkl',
'c:\\Users\\flakej\\Dropbox\\Research\\Data\\CIQ_Transcripts\\2024\\output\\ciq
transcriptcomponent_17.pkl',
'c:\\Users\\flakej\\Dropbox\\Research\\Data\\CIQ_Transcripts\\2024\\output\\ciq
transcriptcomponent_18.pkl',
'c:\\Users\\flakej\\Dropbox\\Research\\Data\\CIQ_Transcripts\\2024\\output\\ciq
transcriptcomponent_19.pkl',
'c:\\Users\\flakej\\Dropbox\\Research\\Data\\CIQ_Transcripts\\2024\\output\\ciq
transcriptcomponent_20.pkl',
'c:\\Users\\flakej\\Dropbox\\Research\\Data\\CIQ_Transcripts\\2024\\output\\ciq
transcriptcomponent_21.pkl',
'c:\\Users\\flakej\\Dropbox\\Research\\Data\\CIQ_Transcripts\\2024\\output\\ciq
transcriptcomponent_22.pkl',
'c:\\Users\\flakej\\Dropbox\\Research\\Data\\CIQ_Transcripts\\2024\\output\\ciq
transcriptcomponent_23.pkl',
'c:\\Users\\flakej\\Dropbox\\Research\\Data\\CIQ_Transcripts\\2024\\output\\ciq
transcriptcomponent_24.pkl',
'c:\\Users\\flakej\\Dropbox\\Research\\Data\\CIQ_Transcripts\\2024\\output\\ciq
transcriptcomponent_25.pkl',
'c:\\Users\\flakej\\Dropbox\\Research\\Data\\CIQ_Transcripts\\2024\\output\\ciq
transcriptcomponent_26.pkl',
'c:\\Users\\flakej\\Dropbox\\Research\\Data\\CIQ_Transcripts\\2024\\output\\ciq
transcriptcomponent_27.pkl',
'c:\\Users\\flakej\\Dropbox\\Research\\Data\\CIQ_Transcripts\\2024\\output\\ciq
transcriptcomponent_28.pkl',
'c:\\Users\\flakej\\Dropbox\\Research\\Data\\CIQ_Transcripts\\2024\\output\\ciq
transcriptcomponent_29.pkl',
'c:\\Users\\flakej\\Dropbox\\Research\\Data\\CIQ_Transcripts\\2024\\output\\ciq
transcriptcomponent_30.pkl']

```

```

[26]: def get_component(filelist):
        df_out = pd.DataFrame()
        for file in filelist:
            print("\n### Processing {} ###".format(file))
            df1 = pd.read_pickle(file, compression="zip")[['transcriptid',
↪ 'transcriptcomponentid']]
            print("Input DF contains {} obs".format(df1.shape[0]))
            df_out = pd.concat([df_out, df1], ignore_index=True)
            print("Aggregated DF contains {} obs".format(df_out.shape[0]))
        return df_out

```

```

[27]: df1 = get_component(filelist[0:15])
        df1.info()

```

```
### Processing c:\Users\flakej\Dropbox\Research\Data\CIQ_Transcripts\2024\output
\ciqtranscriptcomponent_01.pkl ###
Input DF contains 3860671 obs
Aggregated DF contains 3860671 obs
```

```
### Processing c:\Users\flakej\Dropbox\Research\Data\CIQ_Transcripts\2024\output
\ciqtranscriptcomponent_02.pkl ###
Input DF contains 3377437 obs
Aggregated DF contains 7238108 obs
```

```
### Processing c:\Users\flakej\Dropbox\Research\Data\CIQ_Transcripts\2024\output
\ciqtranscriptcomponent_03.pkl ###
Input DF contains 2964202 obs
Aggregated DF contains 10202310 obs
```

```
### Processing c:\Users\flakej\Dropbox\Research\Data\CIQ_Transcripts\2024\output
\ciqtranscriptcomponent_04.pkl ###
```

```
-----
OSError                                Traceback (most recent call last)
Cell In[27], line 1
----> 1 df1 = get_component(filelist[0:15])
      2 df1.info()
```

```
Cell In[26], line 5, in get_component(filelist)
      3 for file in filelist:
      4     print("\n### Processing {} ###".format(file))
----> 5     df1 = pd.read_pickle(file, compression="zip")[['transcriptid',
    ↪ 'transcriptcomponentid']]
      6     print("Input DF contains {} obs".format(df1.shape[0]))
      7     df_out = pd.concat([df_out,df1], ignore_index=True)
```

```
File c:
  ↪ \Users\flakej\AppData\Local\Programs\Python\Python311\Lib\site-packages\pandas\io\pickle.
  ↪ py:179, in read_pickle(filepath_or_buffer, compression, storage_options)
      115 """
      116 Load pickled pandas object (or any object) from file.
      117
      (...)
      176 4      4      9
      177 """
      178 excs_toCatch = (AttributeError, ImportError, ModuleNotFoundError,
    ↪ TypeError)
--> 179 with get_handle(
      180     filepath_or_buffer,
      181     "rb",
      182     compression=compression,
      183     is_text=False,
```

```

184     storage_options=storage_options,
185 ) as handles:
186     # 1) try standard library Pickle
187     # 2) try pickle_compat (older pandas version) to handle subclass
↳changes
188     # 3) try pickle_compat with latin-1 encoding upon a
↳UnicodeDecodeError
190     try:
191         # TypeError for Cython complaints about object.__new__ vs Tick.
↳__new__
192         try:

```

File c:

```

↳\Users\flakej\AppData\Local\Programs\Python\Python311\Lib\site-packages\pandas\io\common.
py:782, in get_handle(path_or_buf, mode, encoding, compression, memory_map,
is_text, errors, storage_options)
777 # ZIP Compression
778 elif compression == "zip":
779     # error: Argument 1 to "_BytesZipFile" has incompatible type
780     # "Union[str, BaseBuffer]"; expected "Union[Union[str,
↳PathLike[str]],
781     # ReadBuffer[bytes], WriteBuffer[bytes]]"
--> 782     handle = _BytesZipFile(
783         handle, ioargs.mode, **compression_args # type: ignore[arg-type]
784     )
785     if handle.buffer.mode == "r":
786         handles.append(handle)

```

File c:

```

↳\Users\flakej\AppData\Local\Programs\Python\Python311\Lib\site-packages\pandas\io\common.
py:1025, in _BytesZipFile.__init__(self, file, mode, archive_name, **kwargs)
1021 kwargs.setdefault("compression", zipfile.ZIP_DEFLATED)
1022 # error: Argument 1 to "ZipFile" has incompatible type "Union[
1023 # Union[str, PathLike[str]], ReadBuffer[bytes], WriteBuffer[bytes]]";
1024 # expected "Union[Union[str, PathLike[str]], IO[bytes]]"
-> 1025 self.buffer = zipfile.ZipFile(file, mode, **kwargs)

```

File c:\Users\flakej\AppData\Local\Programs\Python\Python311\Lib\zipfile.py:

```

↳1284, in ZipFile.__init__(self, file, mode, compression, allowZip64,
compresslevel, strict_timestamps, metadata_encoding)
1282 while True:
1283     try:
-> 1284         self.fp = io.open(file, filemode)
1285     except OSError:
1286         if filemode in modeDict:

```

```

OSError: [Errno 22] Invalid argument: 'c:
↪\\Users\\flakej\\Dropbox\\Research\\Data\\CIQ_Transcripts\\2024\\output\\ciqt_transcriptcomp
↪pkl'

```

```

[62]: df2 = get_component(filelist[15:])
df2.info()

```

```

[62]:      transcriptid  transcriptcomponentid  componentorder
transcriptcomponenttypeid      transcriptcomponenttypename
transcriptpersonid      transcriptpersonname      proid companyofperson
speakertypeid speakertypename \
71218  273,690.0000      13,576,719.0000      51
4      Answer      148,453.0000  Constantine
Karayannopoulos  26,874,765.0000      None      2      Executives
71219  273,690.0000      13,576,720.0000      52
4      Answer      171,086.0000
Mark Smith  102,220,392.0000      None      2      Executives
71220  273,690.0000      13,576,721.0000      53
7  Question and Answer Operator Message      1.0000
Operator      NaN      None      1      Operator
71221  273,690.0000      13,576,722.0000      54
4      Answer      171,086.0000
Mark Smith  102,220,392.0000      None      2      Executives
71222  273,690.0000      13,576,723.0000      55
7  Question and Answer Operator Message      1.0000
Operator      NaN      None      1      Operator

      componenttextpreview  word_count
71218  Well, again as much as I like to agree with yo...      78.0000
71219      Thank you, Constantine.      3.0000
71220  This is all the time you for questions. I woul...      25.0000
71221  Okay. Thanks, operator. I'd like to thank ever...      172.0000
71222  Ladies and gentlemen that concludes today's co...      21.0000

```

Any duplicates at all and at the transcriptid-transcriptcomponenttypeid level?

```

[64]: pd.concat([df1,df2], ignore_index=True).duplicated().value_counts() #confirmed
↪no duplicates

```

```

[64]: False      1541081
Name: count, dtype: int64

```

```

[65]: df3 = pd.
↪concat([df1[['transcriptid','transcriptcomponentid']],df2[['transcriptid','transcriptcompon
↪ignore_index=True)

```

```

<class 'pandas.core.frame.DataFrame'>
Index: 1575627 entries, 0 to 75626

```

Data columns (total 16 columns):

#	Column	Non-Null Count	Dtype
0	companyid	1575627 non-null	object
1	keydevid	1575627 non-null	object
2	transcriptid	1575627 non-null	object
3	headline	1575627 non-null	object
4	mostimportantdateutc	1575627 non-null	object
5	mostimportanttimeutc	1575627 non-null	object
6	keydeventtypeid	1575627 non-null	object
7	keydeventtypename	1575627 non-null	object
8	companyname	1574753 non-null	object
9	transcriptcollectiontypeid	1575627 non-null	int64
10	transcriptcollectiontypename	1575627 non-null	object
11	transcriptpresentationtypeid	1575627 non-null	int64
12	transcriptpresentationtypename	1575627 non-null	object
13	transcriptcreationdate_utc	1575627 non-null	object
14	transcriptcreationtime_utc	1575627 non-null	object
15	audiolengthsec	1532987 non-null	object

dtypes: int64(2), object(14)

memory usage: 1.7 GB

```
[67]: df3.duplicated().value_counts()
```

The previous step shows 2,281 right_only records, inspect them

```
[68]: df3.info()
```

```
[68]: _merge
both          1573346
right_only    2281
left_only      0
Name: count, dtype: int64
```

Compare with wrds_transcript_person

```
[99]: transcript_person = pd.read_pickle("../output/wrds_transcript_person.pkl",
    ↪compression="zip")
```

Verified no duplicates at the transcriptid-transcriptcomponentid level

```
[100]: trans_r_only = right_only['transcriptid'].astype(int).unique().tolist()
len(trans_r_only)
```

```
[100]: 2276
```

Download missing records

```
[80]: import wrds
db = wrds.Connection(wrds_username = '#####')
```

Loading library list...
Done

```
[81]: new = db.raw_sql("select distinct * from ciq_transcripts.wrds_transcript_person_
↳where transcriptid in "+str(tuple(trans_r_only)))
```

```
[108]: new_det = db.raw_sql("select distinct * from ciq_transcripts.
↳wrds_transcript_detail where transcriptid in "+str(tuple(trans_r_only)))
new_det.head()
```

```
[108]:      companyid      keydevid transcriptid
headline mostimportantdateutc mostimportanttimeutc keydeveventtypeid
keydeveventtypename      companyname \
0   873,976.0000 137,253,286.0000 151,041.0000  BNP Paribas, Q2 2011 Earnings
Call, Aug 02, 2011          2011-08-02          13:30:00          48.0000
Earnings Calls          BNP Paribas SA
1   29,279.0000  6,180,480.0000 16,139.0000  Harman International Industries
Inc., Q2 2009 ...          2009-02-04          21:40:00          48.0000
Earnings Calls Harman International Industries, Incorporated
2   36,118.0000  5,979,678.0000 15,106.0000  Geeknet, Inc., Q1 2009 Earnings
Call, Nov-25-2008          2008-11-25          22:00:00          48.0000
Earnings Calls          Geeknet, Inc.
3   354,995.0000 118,231,910.0000 96,055.0000  International Speedway Corp.,
Q4 2010 Earnings...          2011-01-27          14:00:00          48.0000
Earnings Calls          International Speedway Corporation
4  4,169,095.0000  5,606,560.0000 11,510.0000  Triple-S Management
Corporation, Q2 2008 Earni...          2008-08-05          14:00:00
48.0000      Earnings Calls          Triple-S Management Corporation
```

```
      transcriptcollectiontypeid transcriptcollectiontypename
transcriptpresentationtypeid transcriptpresentationtypename
transcriptcreationdate_utc transcriptcreationtime_utc audiolengthsec
0          6          SA Edited Copy
5          Final          2011-08-03
15:02:52      6,533.0000
1          6          SA Edited Copy
5          Final          2009-02-05
14:44:54      NaN
2          6          SA Edited Copy
5          Final          2008-11-26
05:08:05      NaN
3          6          SA Edited Copy
5          Final          2011-01-27
19:25:21      3,171.0000
4          6          SA Edited Copy
5          Final          2008-08-22
21:59:16      NaN
```

```
[113]: df[df.transcriptid.isin([166319,151041])]
```

```
[113]:      transcriptid  saved
109047  166,319.0000      1
```

```
[115]: db.raw_sql("select distinct * from ciq_transcripts.wrds_transcript_detail where
↳transcriptid in "+str(tuple([166319,151041])))
```

```
[115]:      companyid      keydevid transcriptid
headline mostimportantdateutc mostimportanttimeutc keydeveventtypeid
keydeveventtypename      companyname transcriptcollectiontypeid \
0 873,976.0000 137,253,286.0000 151,041.0000 BNP Paribas, Q2 2011 Earnings
Call, Aug 02, 2011      2011-08-02      13:30:00      48.0000
Earnings Calls BNP Paribas SA      6
1 873,976.0000 137,253,286.0000 166,319.0000 BNP Paribas, Q2 2011 Earnings
Call, Aug 02, 2011      2011-08-02      13:30:00      48.0000
Earnings Calls BNP Paribas SA      8
```

```
      transcriptcollectiontypename transcriptpresentationtypeid
transcriptpresentationtypename transcriptcreationdate_utc
transcriptcreationtime_utc audiolengthsec
0      SA Edited Copy      5
Final      2011-08-03      15:02:52      6,533.0000
1      Audited Copy      5
Final      2011-09-08      07:34:34      6,533.0000
```

Same transcript information, but the first id = 151041 is not in the person or full-transcript data set

```
[92]: new = new.merge(right=df2.loc[df2._merge=='right_only',['transcriptid']],
↳how="outer", indicator=True)
```

```
[93]: new._merge.value_counts()
```

```
[93]: _merge
right_only      2280
both              79
left_only         0
Name: count, dtype: int64
```

Of the 2,280 transcripts from detail without corresponding, only one of them has full or person transcript data

```
[94]: new = new.drop(columns=['_merge'])
new.head()
```

```
[94]:      transcriptid transcriptcomponentid componentorder transcriptcomponenttypeid
transcriptcomponenttypename transcriptpersonid transcriptpersonname proid
companyofperson speakertypid speakertypename \
```

0	108.0000	30,233.0000	50.0000	3.0000
Question	684.0000	Ajit Pai	NaN	Thomas Weisel Partners
3.0000	Analysts			
1	108.0000	30,198.0000	15.0000	3.0000
Question	932.0000	Deane Dray	NaN	Goldman Sachs
3.0000	Analysts			
2	108.0000	30,202.0000	19.0000	3.0000
Question	896.0000	Darryl Pardi	NaN	Merrill Lynch
3.0000	Analysts			
3	108.0000	30,244.0000	61.0000	4.0000
Answer	12.0000	Adrian Dillon	NaN	None
2.0000	Executives			
4	108.0000	30,186.0000	3.0000	2.0000
Presenter Speech	650.0000	William P. Sullivan	NaN	
None	2.0000	Executives		

	component	textpreview	word_count
0		Got it. Okay, thank you so much.	7.0000
1		So, this, but there are no other changes in ha...	33.0000
2		Hey, good evening guys.	4.0000
3		I can't do it off the top of my head.	11.0000
4		Thanks, Hilliard and hello everyone. We are p...	776.0000

```
[95]: new.to_pickle("../output/wrds_transcript_person_x.pkl", compression='zip')
```