In [1]:

```
import os
import json
from pathlib import Path
import zipfile
import email
from email.policy import default
from email.parser import Parser
from datetime import timezone
from collections import namedtuple

import pandas as pd
import s3fs
from bs4 import BeautifulSoup
from dateutil.parser import parse
from chardet.universaldetector import UniversalDetector

import pyspark
from pyspark.ml import Pipeline
from pyspark.ml.feature import CountVectorizer
from pyspark.ml.feature import HashingTF, Tokenizer
from pyspark.sql import SparkSession
from pyspark.sql.functions import col
from pyspark.ml.pipeline import Transformer
from pyspark.sql.functions import udf
from pyspark.sql.types import StructType, StringType,StructField

import pandas as pd
```

In [2]:

```
current_dir = Path(os.getcwd()).absolute()
results_dir = current_dir.joinpath('results')
results_dir.mkdir(parents=True, exist_ok=True)
data_dir = current_dir.joinpath('data')
data_dir.mkdir(parents=True, exist_ok=True)
enron_data_dir = data_dir.joinpath('enron')


output_columns = [
        'payload',
        'text',
        'Message_D',
        'Date',
        'From',
        'To',
        'Subject',
        'Mime-Version',
        'Content-Type',
        'Content-Transfer-Encoding',
        'X-From',
        'X-To',
        'X-cc',
        'X-bcc',
        'X-Folder',
        'X-Origin',
        'X-FileName',
        'Cc',
        'Bcc'
]

columns = [column.replace('-', '_') for column in output_columns]

ParsedEmail = namedtuple('ParsedEmail', columns)

spark = SparkSession\
    .builder\
    .appName("Assignment04")\
    .getOrCreate()
```

In [3]:

```
def copy_data_to_local():
    endpoint_url='C:\\Users\\theoj\\Downloads\\Week4\\'


#     s3 = s3fs.S3FileSystem(
#         anon=True,
#         client_kwargs={
#             'endpoint_url': endpoint_url
#         }
#     )

    enron_data_path = 'enron.zip'
    with zipfile.ZipFile(enron_data_path) as f_zip:
      f_zip.extractall(path=data_dir)

copy_data_to_local()
```

Assignment 4.1

In [4]:

```
def read_raw_email(email_path):
    detector = UniversalDetector()

    try:
        with open(email_path) as f:
            original_msg = f.read()
    except UnicodeDecodeError:
        detector.reset()
        with open(email_path, 'rb') as f:
            for line in f.readlines():
                detector.feed(line)
                if detector.done:
                    break
        detector.close()
        encoding = detector.result['encoding']
        with open(email_path, encoding=encoding) as f:
            original_msg = f.read()

    return original_msg

def make_spark_df():
    records = []
    #print(enron_data_dir)
    for root, dirs, files in os.walk(enron_data_dir):
#         print ("files")
        for file_path in files:
            ## Current path is now the file path to the current email.
            ## Use this path to read the following information
            ## original_msg
            ## username (Hint: It is the root folder)
            ## id (The relative path of the email message)
            current_path = Path(root).joinpath(file_path)
            record = {}
            username = os.path.basename(os.path.dirname(root))
            id = username+"/"+os.path.basename(root)+"/"+file_path

            print(id,username,read_raw_email(current_path))
#             record["username"]=username
            test = read_raw_email(current_path)
            email_id = test[(test.find('Message-ID: <')+13):(test.find('>\nDate:
            email_from = test[(test.find('\nFrom: ')+7):(test.find('\nTo:'))]
            email_to = test[(test.find('\nTo: ')+5):(test.find('\nSubject: '))]
            email_subject = test[(test.find('\nSubject: ')+10):(test.find('\nMime
            record["id"]=id
            record["username"]=username
            record["email_id"]=email_id
            record["from"]=email_from
            record["to"]=email_to
            record["subject"]=email_subject
            record["original_msg"]=test[:350]+test[-150:]
            record["email_path"]=file_path
            records.append(record)
#             print(test[(test.find('\nSubject: ')+10):(test.find('\nMime-Versio
    ## TODO: Complete the code to code to create the Spark dataframe

    schemaString = "id email_id username from to subject original_msg email_path"
```

```
#     StructType([ \
#     StructField("id",StringType(),True), \
#     StructField("email_id",StringType(),True), \
#     StructField("username",StringType(),True), \
#     StructField("from",StringType(),True), \
#     StructField("to",StringType(),True), \
#     StructField("subject",StringType(),True), \
#     StructField("message", StringType(), True) \
#   ])
# "id from to subject original_msg" schemaString.split()
    fields = [StructField(field_name,StringType(),True) for field_name in schemaS
    schema = StructType(fields)
#     print(record)
    return spark.createDataFrame(records, schema)

#     return spark.createDataFrame(data=records,schema = schema1)
# df2.printSchema()
# df2.show(truncate=False)

df = make_spark_df()
```

```
meyers-a/deleted_items/686_ meyers-a Message-ID: <29402724.1075841306157.Java
Mail.evans@thyme>
Date: Mon, 14 Jan 2002 03:36:42 -0800 (PST)
From: pete.davis@enron.com
To: pete.davis@enron.com
Subject: Start Date: 1/14/02; HourAhead hour: 6;
Cc: albert.meyers@enron.com, bill.williams@enron.com, craig.dean@enron.com,
        geir.solberg@enron.com, john.anderson@enron.com,
        mark.guzman@enron.com, michael.mier@enron.com, pete.davis@enron.com,
        ryan.slinger@enron.com
Mime-Version: 1.0
Content-Type: text/plain; charset=us-ascii
Content-Transfer-Encoding: 7bit
Bcc: albert.meyers@enron.com, bill.williams@enron.com, craig.dean@enron.com,
        geir.solberg@enron.com, john.anderson@enron.com,
        mark.guzman@enron.com, michael.mier@enron.com, pete.davis@enron.com,
        ryan.slinger@enron.com
X-From: Davis, Pete </O=ENRON/OU=NA/CN=RECIPIENTS/CN=PDAVIS1>
X-To: Davis, Pete </O=ENRON/OU=NA/CN=RECIPIENTS/CN=PDAVIS1>
```

In [5]:
```
df.show()
```

```
+--------------------+--------------------+--------+--------------------+------
--------------+--------------------+--------------------+----------+
|                  id|            email_id|username|                from|
to|             subject|        original_msg|email_path|
+--------------------+--------------------+--------+--------------------+------
--------------+--------------------+--------------------+----------+
|meyers-a/deleted_...|29402724.10758413...|meyers-a|pete.davis@enron.com|pete.d
avis@enron.com|Start Date: 1/14/...|Message-ID: <2940...|      686_|
|meyers-a/deleted_...|12395403.10758413...|meyers-a|pete.davis@enron.com|pete.d
avis@enron.com|Start Date: 1/6/0...|Message-ID: <1239...|      889_|
|meyers-a/deleted_...|23700826.10758412...|meyers-a|pete.davis@enron.com|pete.d
avis@enron.com|Start Date: 1/28/...|Message-ID: <2370...|      242_|
|meyers-a/deleted_...|10294441.10758413...|meyers-a|pete.davis@enron.com|pete.d
avis@enron.com|Start Date: 1/7/0...|Message-ID: <1029...|      880_|
|meyers-a/deleted_...|7932327.107584130...|meyers-a|pete.davis@enron.com|pete.d
avis@enron.com|Start Date: 1/10/...|Message-ID: <7932...|      782_|
|meyers-a/deleted_...|2679417.107584131...|meyers-a|pete.davis@enron.com|pete.d
avis@enron.com|Start Date: 1/2/0...|Message-ID: <2679...|     1018_|
|meyers-a/deleted_...|32879300.10758412...|meyers-a|no.address@enron....|age-I
D: <32879300...|Copier Commitment...|Message-ID: <3287...|      377_|
|meyers-a/deleted_...|32748140.10758413...|meyers-a|bert.meyers@enron...|leaf.h
arasin@enro...|TAG 25883
Cc: bil...|Message-ID: <3274...|     1120_|
|meyers-a/deleted_...|21654541.10758413...|meyers-a|pete.davis@enron.com|pete.d
avis@enron.com|Schedule Crawler:...|Message-ID: <2165...|      578_|
|meyers-a/deleted_...|7613902.107584130...|meyers-a|pete.davis@enron.com|pete.d
avis@enron.com|Schedule Crawler:...|Message-ID: <7613...|      498_|
|meyers-a/deleted_...|24318539.10758413...|meyers-a|bert.meyers@enron...|ryan.s
linger@enro...|Dates to Keep in ...|Message-ID: <2431...|     1124_|
|meyers-a/deleted_...|5118583.107584131...|meyers-a|pete.davis@enron.com|pete.d
avis@enron.com|Start Date: 1/3/0...|Message-ID: <5118...|      980_|
|meyers-a/deleted_...|15179708.10758413...|meyers-a|pete.davis@enron.com|pete.d
avis@enron.com|Schedule Crawler:...|Message-ID: <1517...|      572_|
|meyers-a/deleted_...|12311986.10758412...|meyers-a|pete.davis@enron.com|pete.d
avis@enron.com|Start Date: 1/30/...|Message-ID: <1231...|      196_|
|meyers-a/deleted_...|9915376.107584129...|meyers-a|pete.davis@enron.com|pete.d
avis@enron.com|Start Date: 1/29/...|Message-ID: <9915...|      210_|
|meyers-a/deleted_...|26417709.10758412...|meyers-a|pete.davis@enron.com|pete.d
avis@enron.com|Schedule Crawler:...|Message-ID: <2641...|      358_|
|meyers-a/deleted_...|18733926.10758412...|meyers-a|pete.davis@enron.com|pete.d
avis@enron.com|Start Date: 1/26/...|Message-ID: <1873...|      308_|
|meyers-a/deleted_...|19046109.10758413...|meyers-a|pete.davis@enron.com|pete.d
avis@enron.com|Start Date: 1/2/0...|Message-ID: <1904...|     1004_|
|meyers-a/deleted_...|12451678.10758413...|meyers-a|pete.davis@enron.com|pete.d
avis@enron.com|Start Date: 1/11/...|Message-ID: <1245...|      752_|
|meyers-a/deleted_...|24225764.10758413...|meyers-a|pete.davis@enron.com|pete.d
avis@enron.com|Start Date: 1/14/...|Message-ID: <2422...|      672_|
+--------------------+--------------------+--------+--------------------+------
--------------+--------------------+--------------------+----------+
only showing top 20 rows
```

In [6]:

```
df.printSchema()
```

```
root
 |-- id: string (nullable = true)
 |-- email_id: string (nullable = true)
 |-- username: string (nullable = true)
 |-- from: string (nullable = true)
 |-- to: string (nullable = true)
 |-- subject: string (nullable = true)
 |-- original_msg: string (nullable = true)
 |-- email_path: string (nullable = true)
```

Assignment 4.2

In [7]:

```
plain_msg_example = """
Message-ID: <6742786.1075845426893.JavaMail.evans@thyme>
Date: Thu, 7 Jun 2001 11:05:33 -0700 (PDT)
From: jeffrey.hammad@enron.com
To: andy.zipper@enron.com
Subject: Thanks for the interview
Mime-Version: 1.0
Content-Type: text/plain; charset=us-ascii
Content-Transfer-Encoding: 7bit
X-From: Hammad, Jeffrey </O=ENRON/OU=NA/CN=RECIPIENTS/CN=NOTESADDR/CN=CBBE377A-24
X-To: Zipper, Andy </O=ENRON/OU=NA/CN=RECIPIENTS/CN=AZIPPER>
X-cc: 
X-bcc: 
X-Folder: \Zipper, Andy\Zipper, Andy\Inbox
X-Origin: ZIPPER-A
X-FileName: Zipper, Andy.pst

Andy,

Thanks for giving me the opportunity to meet with you about the Analyst/ Associat

Thanks and Best Regards,

Jeff Hammad
"""

html_msg_example = """
Message-ID: <21013632.1075862392611.JavaMail.evans@thyme>
Date: Mon, 19 Nov 2001 12:15:44 -0800 (PST)
From: insynconline.6jy5ympb.d@insync-palm.com
To: tstaab@enron.com
Subject: Last chance for special offer on Palm OS Upgrade!
Mime-Version: 1.0
Content-Type: text/plain; charset=us-ascii
Content-Transfer-Encoding: 7bit
X-From: InSync Online <InSyncOnline.6jy5ympb.d@insync-palm.com>
X-To: THERESA STAAB <tstaab@enron.com>
X-cc: 
X-bcc: 
X-Folder: \TSTAAB (Non-Privileged)\Staab, Theresa\Deleted Items
X-Origin: Staab-T
X-FileName: TSTAAB (Non-Privileged).pst

<html>

<html>
<head>
<title>Paprika</title>
<meta http-equiv="Content-Type" content="text/html;">
</head>
<body bgcolor="#FFFFFF" TEXT="#333333" LINK="#336699" VLINK="#6699cc" ALINK="#ff9
<table border="0" cellpadding="0" cellspacing="0" width="582">
<tr valign="top">
  <td width="582" colspan="9"><nobr><a href="http://insync-online.p04.com/u.d?BEF
</tr>
<tr valign="top">
```

```
  <td width="4" bgcolor="#CCCCCC"><img src="http://images4.postdirect.com/master-
  <td width="20"><img src="http://images4.postdirect.com/master-images/404707/cle
  <td width="165"><br><a href="http://insync-online.p04.com/u.d?LkReaQA5eczXL=21'
  <td width="20"><img src="http://images4.postdirect.com/master-images/404707/cle
  <td width="165"><br><a href="http://insync-online.p04.com/u.d?BkReaQA5eczXO=31'
  <td width="20"><img src="http://images4.postdirect.com/master-images/404707/cle
  <td width="165"><br><a href="http://insync-online.p04.com/u.d?JkReaQA5eczXRs=41
  <td width="19"><img src="http://images4.postdirect.com/master-images/404707/cle
  <td width="4" bgcolor="#CCCCCC"><img src="http://images4.postdirect.com/master-
</tr>
</table>
<table border="0" cellpadding="0" cellspacing="0" width="582">
<tr valign="top">
  <td width="4" bgcolor="#CCCCCC"><img src="http://images4.postdirect.com/master-
  <td width="574"><br>
    <table border="0" cellpadding="0" cellspacing="0" width="574" bgcolor="#99ccf
    <tr>
      <td width="50"><img src="http://images4.postdirect.com/master-images/404707
      <td width="474"><font face="verdana, arial" size="-2"color="#000000">
        <br>
        Dear THERESA,
        <br><br>
        Due to overwhelming demand for the Palm OS&#174; v4.1 Upgrade with Mobile
        extending the special offer of 25% off through November 30, 2001. So ther
        increase the functionality of your Palm&#153; III, IIIx, IIIxe, IIIc, V c
        new Palm OS v4.1 through this extended special offer. You'll receive the 
        <b>for just $29.95 when you use Promo Code <font color="#FF0000">OS41WAVE
        <b>$10 savings</b> off the list price.
        <br><br>
        <a href="http://insync-online.p04.com/u.d?NkReaQA5eczXRh=51">Click here t
        <br><br>
        <a href="http://insync-online.p04.com/u.d?MkReaQA5eczXRm=61"><img src="ht
        <br><br>
        You can do a lot more with your Palm&#153; handheld when you upgrade to t
        favorite features just got even better and there are some terrific new ad
        <br><br>
        <LI> Handwrite notes and even draw pictures right on your Palm&#153 handh
        <LI> Tap letters with your stylus and use Graffiti&#174; at the same time
        <LI> Improved Date Book functionality lets you view, snooze or clear mult
        <LI> You can easily change time-zone settings</LI>

        <br><br>
        <a href="http://insync-online.p04.com/u.d?WkReaQA5eczXRb=71"><img src="ht
        <br><br>
        <LI> <nobr>Mask/unmask</nobr> private records or hide/unhide directly wit
        <LI> Lock your device automatically at a designated time using the new Au
        <LI> Always remember your password with our new Hint feature*</LI>

        <br><br>
        <a href="http://insync-online.p04.com/u.d?VEReaQA5eczXRQ=81"><img src="ht
        <br><br>
        <LI> Use your GSM compatible mobile phone or modem to get online and acce
        <LI> Stay connected with email, instant messaging and text messaging to G
        <LI> Send applications or records through your cell phone to schedule mee
             important information to others</LI>

        <br><br>
        All this comes in a new operating system that can be yours for just $29.9
        upgrade to the new Palm&#153; OS v4.1</a> and you'll also get the latest
        <nobr>1-800-881-7256</nobr> to order via phone.
        <br><br>
        Sincerely,<br>
        The Palm Team
        <br><br>
        P.S. Remember, this extended offer opportunity of 25% savings absolutely
        and is only available through the Palm Store when you use Promo Code <b><
        <br><br>
        <img src="http://images4.postdirect.com/master-images/404707/bottom_butto
        <br><img src="http://images4.postdirect.com/master-images/404707/clear.gi
        </font></td>
      <td width="50"><img src="http://images4.postdirect.com/master-images/404707
    </tr>
    </table></td>
    <td width="4" bgcolor="#CCCCCC"><img src="http://images4.postdirect.com/maste
  </tr>
  <tr>
  <td colspan="3"><img src="http://images4.postdirect.com/master-images/404707/bo
  </tr>
</table>
<table border="0" cellpadding="0" cellspacing="0" width="582">
  <tr>
    <td width="54"><img src="http://images4.postdirect.com/master-images/404707/c
    <td width="474"><font face="arial, verdana" size="-2" color="#000000"><br>
    * This feature is available on the Palm&#153; IIIx, Palm&#153; IIIxe, and Pal
    ** Note: To use the MIK functionality, you need either a Palm OS&#174; compat
    with  <nobr>built-in</nobr> modem or data capability that has either an infra
    are using a phone, you must have data services from your mobile service provi
    a list of tested and supported phones that you can use with the MIK. Cable no
    <br><br>
    ------------------<br>
    To modify your profile or unsubscribe from Palm newsletters, <a href="http://
    Or, unsubscribe by replying to this message, with "unsubscribe" as the subjec
    <br><br>
    ------------------<br>
    Copyright&#169; 2001 Palm, Inc. Palm OS, Palm Computing, HandFAX, HandSTAMP,
    HotSync, iMessenger, MultiMail, Palm.Net, PalmConnect, PalmGlove, PalmModem,
    and the Palm Platform Compatible Logo are registered trademarks of Palm, Inc.
    AnyDay, EventClub, HandMAIL, the HotSync Logo, PalmGear, PalmGlove, PalmPix,
    trade dress, PalmSource, Smartcode, and Simply Palm are trademarks of Palm, I
    product names may be trademarks or registered trademarks of their respective
    <img src="http://images4.postdirect.com/master-images/404707/clear.gif" width
    <td width="54"><img src="http://images4.postdirect.com/master-images/404707/c
  </tr>
</table><br><br><br><br>
<!-- The following image is included for message detection -->
<img src="http://p04.com/1x1.dyn" border="0" alt="" width="1" height="1">
<img src="http://p04.com/1x1.dyn?0vEGou8Ig30ba2L2bLn" width=1 height=1></body>
</html>

</html>
"""
plain_msg_example = plain_msg_example.strip()
html_msg_example = html_msg_example.strip()
print(plain_msg_example)
```

```
Message-ID: <6742786.1075845426893.JavaMail.evans@thyme>
Date: Thu, 7 Jun 2001 11:05:33 -0700 (PDT)
From: jeffrey.hammad@enron.com
To: andy.zipper@enron.com
Subject: Thanks for the interview
Mime-Version: 1.0
Content-Type: text/plain; charset=us-ascii
Content-Transfer-Encoding: 7bit
X-From: Hammad, Jeffrey </O=ENRON/OU=NA/CN=RECIPIENTS/CN=NOTESADDR/CN=CBBE377A-
24F58854-862567DD-591AE7>
X-To: Zipper, Andy </O=ENRON/OU=NA/CN=RECIPIENTS/CN=AZIPPER>
X-cc:
X-bcc:
X-Folder: \Zipper, Andy\Zipper, Andy\Inbox
X-Origin: ZIPPER-A
X-FileName: Zipper, Andy.pst

Andy,

Thanks for giving me the opportunity to meet with you about the Analyst/ Associ
ate program.  I enjoyed talking to you, and look forward to contributing to the
success that the program has enjoyed.

Thanks and Best Regards,

Jeff Hammad
```

In [8]:

```python
def parse_html_payload(payload):
    """
    This function uses Beautiful Soup to read HTML data
    and return the text.  If the payload is plain text, then
    Beautiful Soup will return the original content
    """
    soup = BeautifulSoup(payload, 'html.parser')
    return str(soup.get_text()).encode('utf-8').decode('utf-8')

def parse_email(original_msg):
    result = {}
    msg = Parser(policy=default).parsestr(original_msg)
    result['payload'] = msg.get_payload()
    result['text'] = parse_html_payload(result['payload'])
    try:
      for key, value in msg.items():
        result[key.replace('-','-')] = value
    except Exception as e:
      print('Problem parsing email: {}\n{}'.format(email_path,e))
    try:
      result['Date'] = parse(result['Date'],ignoretz=False).isoformat()
    except Exception as e:
      print('Problem converting date: {}\n{}'.format(result.get('date'),e))
    tuple_result = tuple([str(result.get(column,None))for column in columns])
    return ParsedEmail(*tuple_result)
```

In [9]:

```python
parsed_msg = parse_email(plain_msg_example)
```

In [10]:

```
print(parsed_msg.text)
```

```
Andy,

Thanks for giving me the opportunity to meet with you about the Analyst/ Associ
ate program.  I enjoyed talking to you, and look forward to contributing to the
success that the program has enjoyed.

Thanks and Best Regards,

Jeff Hammad
```

In [11]:

```
parsed_html_msg =parse_email(html_msg_example)
```

Assignment 4.3

In [12]:

```
## This creates a schema for the email data
email_struct = StructType()

for column in columns:
    email_struct.add(column, StringType(), True)
```

In [13]:

```
## This creates a user-defined function which can be used in Spark
parse_email_func = udf(lambda z: parse_email(z), email_struct)

def parse_emails(input_df):
    new_df = input_df.select(
        'username', 'id', 'original_msg', parse_email_func('original_msg').alias(
    )
    for column in columns:
        new_df = new_df.withColumn(column, new_df.parsed_email[column])

    new_df = new_df.drop('parsed_email')
    return new_df

class ParseEmailsTransformer(Transformer):
    def _transform(self, dataset):
        """
        Transforms the input dataset.

        :param dataset: input dataset, which is an instance of :py:class:`pyspark
        :returns: transformed dataset
        """
        return dataset.transform(parse_emails)

## Use the custom ParseEmailsTransformer, Tokenizer, and CountVectorizer
## to create a spark pipeline
## TODO:  Complete code

parseemailtransformer = ParseEmailsTransformer()
tokenizer = Tokenizer(inputCol="text",outputCol="words")
cv = CountVectorizer(inputCol=tokenizer.getOutputCol(),outputCol="features")
email_pipeline = Pipeline(stages=[parseemailtransformer,tokenizer,cv])
model = email_pipeline.fit(df)
result = model.transform(df)
```

In [14]:
```
result.select('id','words','features').show()
```

```
+--------------------+--------------------+--------------------+
|                  id|               words|            features|
+--------------------+--------------------+--------------------+
|meyers-a/deleted_...|[, , , , log, mes...|(17092,[0,41,42,4...|
|meyers-a/deleted_...|[, , , , log, mes...|(17092,[0,41,42,4...|
|meyers-a/deleted_...|[, , , , log, mes...|(17092,[0,41,42,4...|
|meyers-a/deleted_...|                  []|   (17092,[0],[1.0])|
|meyers-a/deleted_...|[, , , , log, mes...|(17092,[0,41,42,4...|
|meyers-a/deleted_...|[, , trans_type:,...|(17092,[0,42,320,...|
|meyers-a/deleted_...|[, non-bankrupt, ...|(17092,[0,13,20,2...|
|meyers-a/deleted_...|      [bert, meyers]|(17092,[337,593],...|
|meyers-a/deleted_...|[!!!unknown, data...|(17092,[9,10,17,1...|
|meyers-a/deleted_...|[, , start, date:...|(17092,[0,96,116,...|
|meyers-a/deleted_...|[give, me, a, cal...|(17092,[0,9,14,20...|
|meyers-a/deleted_...|[, , , , log, mes...|(17092,[0,41,42,4...|
|meyers-a/deleted_...|[, , start, date:...|(17092,[0,96,116,...|
|meyers-a/deleted_...|[, , , , log, mes...|(17092,[0,41,42,4...|
|meyers-a/deleted_...|[, , , , log, mes...|(17092,[0,41,42,4...|
|meyers-a/deleted_...|[, , start, date:...|(17092,[0,96,116,...|
|meyers-a/deleted_...|[, , , , log, mes...|(17092,[0,41,42,4...|
|meyers-a/deleted_...|[, , , , log, mes...|(17092,[0,41,42,4...|
|meyers-a/deleted_...|[, , , , log, mes...|(17092,[0,41,42,4...|
|meyers-a/deleted_...|[, , , , log, mes...|(17092,[0,41,42,4...|
+--------------------+--------------------+--------------------+
only showing top 20 rows
```

In [15]:
```
# !pip install pyspark
```