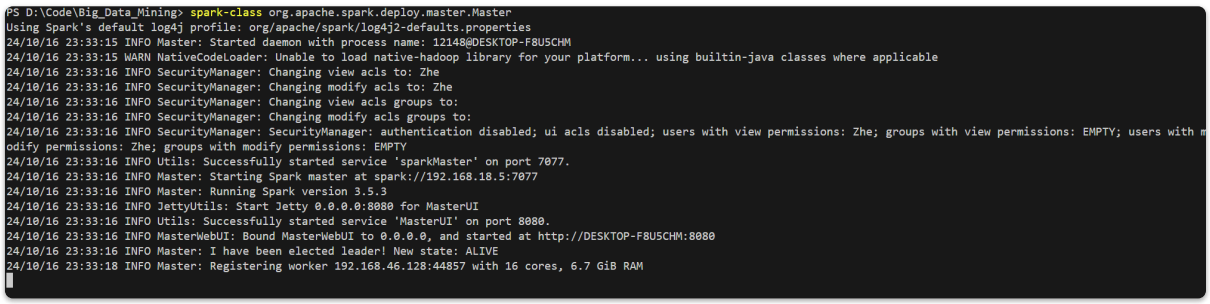
資工碩一 113598007 楊明哲

環境設置

1. Windows
    - 環境版本
        - Java : 17.0.12
        - Python : 3.10.10
        - PySpark : 3.5.3
        - Pandas : 2.2.3
    - 執行指令 spark-class org.apache.spark.deploy.master.Master，並打開 URL
        - 指令結果



2. Ubuntu
    - 環境版本
        - Memory : 8GB
        - Processors : 16
        - Java : 17.0.12
        - Python : 3.10.12
        - Spark : 3.5.3
    - 執行 spark-class org.apache.spark.deply.worker.Worker spark://192.168.18.5:7077

資料夾結構

- homework1.py
- archive
    - spacenews.csv
- image
    - all of output snapshots

源代碼

```python
import pandas as pd
from pyspark.sql import SparkSession
from pyspark.sql.functions import explode, split, col, lower, regexp_replace, count, to_date, date_format

# Create SparkSession
spark = SparkSession.builder.appName("Word Frequency").getOrCreate()

# Read Data
file_path = './archive/spacenews.csv'
df = pd.read_csv(file_path)
df.dropna(inplace = True)

# Convert Pandas DataFrame to Spark DataFrame
spark_df_origin = spark.createDataFrame(df)
spark_df_origin = spark_df_origin.withColumn('Date', to_date(col('Date'), 'MMMM d, yyyy'))
spark_df = spark_df_origin

# Q1 : Total word frequency
spark_df = spark_df.withColumn('Title', lower(col('Title')))
spark_df = spark_df.withColumn('Title', regexp_replace(col('Title'), '[^\w\s]', ''))
spark_df = spark_df.withColumn('word', explode(split(col('Title'), '\s+')))

total_word_freq = spark_df.groupBy('word').count().orderBy(col('count').desc())

total_word_freq = spark_df.groupBy('word').count().orderBy(col('count').desc())
daily_word_freq = spark_df.groupBy('Date', 'word').count().orderBy(col('Date'), col('count').desc())

total_word_freq.show()
daily_word_freq.show()

# Q2 : Total content word frequency
spark_df = spark_df_origin
spark_df = spark_df.withColumn('Content', lower(col('Content')))
spark_df = spark_df.withColumn('Content', regexp_replace(col('Content'), '[^\w\s]', ''))
spark_df = spark_df.withColumn('content_word', explode(split(col('Content'), '\s+')))

total_content_word_freq = spark_df.groupBy('content_word').count().orderBy(col('count').desc())
daily_content_word_freq = spark_df.groupBy('Date', 'content_word').count().orderBy(col('Date'), col('count').desc())

total_content_word_freq.show()
daily_content_word_freq.show()

# Q3 : Daily article count
spark_df = spark_df_origin
daily_article_count = spark_df.groupBy('Date').count().withColumnRenamed('count', 'daily_count')
```

```python
daily_article_percentage = daily_article_count.withColumn('percentage', (col('daily_count') / total_article_count) * 100)

author_daily_article_count = spark_df.groupBy('Date', 'Author').count().withColumnRenamed('count', 'author_daily_count')
author_daily_article_percentage = author_daily_article_count.join(daily_article_count, 'Date')
author_daily_article_percentage = author_daily_article_percentage.withColumn('percentage', (col('author_daily_count') / col('daily_count')) * 100)

daily_article_percentage.show()
author_daily_article_percentage.show()

# Q4 : Space in Title and Postexcerpt
spark_df = spark_df_origin

spark_df = spark_df.withColumn('Title', lower(col('Title')))
spark_df = spark_df.withColumn('Postexcerpt', lower(col('Postexcerpt')))
spark_df = spark_df.withColumn('Title', regexp_replace(col('Title'), '[^\w\s]', ''))
spark_df = spark_df.withColumn('Postexcerpt', regexp_replace(col('Postexcerpt'), '[^\w\s]', ''))

filtered_df = spark_df.filter((col('Title').contains('space')) & (col('Postexcerpt').contains('space')))

sorted_filtered_df = filtered_df.orderBy('Date')

sorted_filtered_df.show()


```

輸出結果

- 執行 spark-submit --master spark://192.168.18.5:7077 --conf spark.driver.host=192.168.18.5 homework1.py

第一題

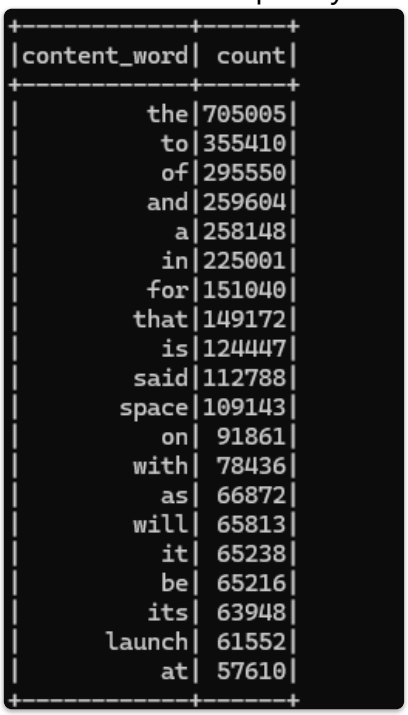
- Title word frequency

```
+----------+-----+
|      word|count|
+----------+-----+
|        to| 5786|
|     space| 4113|
|       for| 3980|
|        of| 2254|
|        in| 2208|
|    launch| 2084|
| satellite| 2024|
|        on| 1922|
|      nasa| 1887|
|       the| 1427|
|     force| 1286|
|        us| 1249|
|       and| 1242|
|      with| 1090|
|         a|  997|
|       new|  984|
|satellites|  901|
|commercial|  851|
|    spacex|  837|
|       air|  780|
+----------+-----+
```

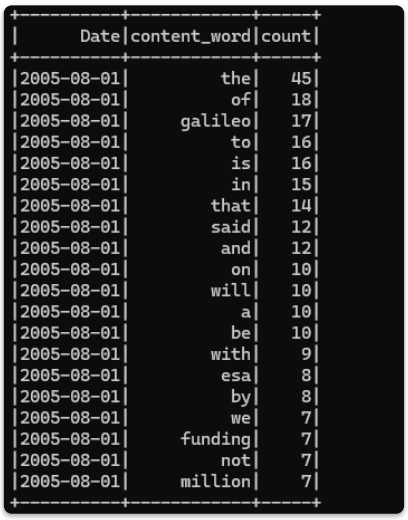
```
+----------+----------+-----+
|      Date|      word|count|
+----------+----------+-----+
|2005-08-01|        in|    1|
|2005-08-01|       esa|    1|
|2005-08-01|      over|    1|
|2005-08-01|       and|    1|
|2005-08-01|    member|    1|
|2005-08-01|  standoff|    1|
|2005-08-01|    states|    1|
|2005-08-01|   galileo|    1|
|2005-11-07|  asteroid|    1|
|2005-11-07|      plan|    1|
|2005-11-07|        to|    1|
|2005-11-07|  possible|    1|
|2005-11-07|    strike|    1|
|2005-11-07|      nasa|    1|
|2005-11-07|       for|    1|
|2005-11-07|     urges|    1|
|2005-11-07|     group|    1|
|2005-11-21|commercial|    1|
|2005-11-21|     cargo|    1|
|2005-11-21|        to|    1|
+----------+----------+-----+
```

第二題

- Content word frequency



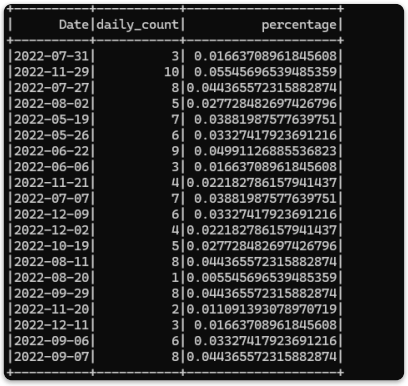
content_word	count
the	705005
to	355410
of	295550
and	259604
a	258148
in	225001
for	151040
that	149172
is	124447
said	112788
space	109143
on	91861
with	78436
as	66872
will	65813
it	65238
be	65216
its	63948
launch	61552
at	57610



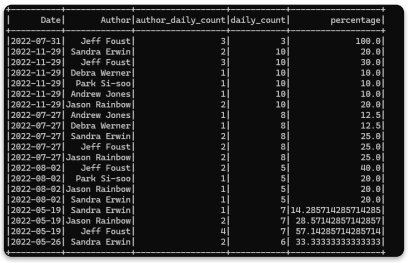
Date	content_word	count
2005-08-01	the	45
2005-08-01	of	18
2005-08-01	galileo	17
2005-08-01	to	16
2005-08-01	is	16
2005-08-01	in	15
2005-08-01	that	14
2005-08-01	said	12
2005-08-01	and	12
2005-08-01	on	10
2005-08-01	will	10
2005-08-01	a	10
2005-08-01	be	10
2005-08-01	with	9
2005-08-01	esa	8
2005-08-01	by	8
2005-08-01	we	7
2005-08-01	funding	7
2005-08-01	not	7
2005-08-01	million	7

第三題

- Daily article count



Date	daily_count	percentage
2022-07-31	3	0.01663708961845608
2022-11-29	10	0.05545696539485359
2022-07-27	8	0.04436557231588274
2022-08-02	5	0.02772848269742796
2022-05-19	7	0.03881987577639751
2022-05-26	6	0.03327417923691216
2022-06-22	9	0.04991126885536823
2022-06-06	3	0.01663708961845608
2022-11-21	4	0.02218278615794143
2022-07-07	7	0.03881987577639751
2022-12-09	6	0.03327417923691216
2022-12-02	4	0.02218278615794143
2022-10-19	5	0.02772848269742796
2022-08-11	8	0.04436557231588274
2022-08-20	1	0.00554569653948536
2022-09-29	8	0.04436557231588274
2022-11-20	2	0.01109139307897071
2022-12-11	3	0.01663708961845608
2022-09-06	6	0.03327417923691216
2022-09-07	8	0.04436557231588274



Date	Author	author_daily_count	daily_count	percentage
2022-07-31	Jeff Foust	3	3	100.0
2022-11-29	Sandra Erwin	2	10	20.0
2022-11-29	Jeff Foust	3	10	30.0
2022-11-29	Debra Werner	1	10	10.0
2022-11-29	Park Si-soo	1	10	10.0
2022-11-29	Andrew Jones	1	10	10.0
2022-11-29	Jason Rainbow	2	10	20.0
2022-07-27	Andrew Jones	1	8	12.5
2022-07-27	Debra Werner	1	8	12.5
2022-07-27	Sandra Erwin	2	8	25.0
2022-07-27	Jeff Foust	2	8	25.0
2022-07-27	Jason Rainbow	2	8	25.0
2022-08-02	Jeff Foust	2	5	40.0
2022-08-02	Park Si-soo	1	5	20.0
2022-08-02	Jason Rainbow	1	5	20.0
2022-08-02	Sandra Erwin	1	5	20.0
2022-05-19	Sandra Erwin	1	7	14.285714285714285
2022-05-19	Jason Rainbow	2	7	28.571428571428573
2022-05-19	Jeff Foust	4	7	57.142857142857144
2022-05-26	Sandra Erwin	2	6	33.33333333333333

第四題

- Space in Title and Postexcerpt

Title	url	content	author	Date	Postexcerpt
despite the budge...	https://spacenews...	WASHINGTON — As t...	Brian Berger	2006-08-29	as the new direct...
lightfoot to lead...	https://spacenews...	WASHINGTON — NASA...	Amy Klamper	2009-08-24	washington nasa ...
mike gold corpora...	https://spacenews...	With the administ...	Amy Klamper	2009-08-31	with the administ...
peckham leaves sp...	https://spacenews...	WASHINGTON — Rob ...	Warren Ferster	2009-09-02	washington rob p...
spacex lands orbc...	https://spacenews...	WASHINGTON — Sate...	Warren Ferster	2009-09-03	washington satel...
ariane 5 remnant ...	https://spacenews...	The remains of a ...	SpaceNews Staff	2009-09-07	the remains of a ...
spacex hardware d...	https://spacenews...	Space Exploration...	SpaceNews Staff	2009-09-07	space exploration...
amid shift in pow...	https://spacenews...	TOKYO — The Japan...	Paul Kallender-Umezu	2009-09-07	tokyo the japane...
japans first spac...	https://spacenews...	NEW YORK — Japan'...	SpaceNews Staff	2009-09-07	new york japans ...
spacex and astriu...	https://spacenews...	WASHINGTON — Hawt...	Amy Klamper	2009-09-08	washington hawth...
white house hande...	https://spacenews...	WASHINGTON — Add ...	Amy Klamper	2009-09-11	washington add m...
us canada sign sp...	https://spacenews...	NASA Administrato...	SpaceNews Staff	2009-09-14	nasa administrato...
arianespace chief...	https://spacenews...	PARIS — Engineers...	Peter B. de Selding	2009-09-14	paris engineers ...
esa designing doc...	https://spacenews...	PARIS — The Europ...	Peter B. de Selding	2009-09-18	paris the europe...
us france sign fo...	https://spacenews...	NASA Administrato...	SpaceNews Staff	2009-09-21	nasa administrato...
a 50year history ...	https://spacenews...	BERKELEY, Calif. ...	Debra Werner	2009-09-21	berkeley calif a...
volker liebig dir...	https://spacenews...	Satellite Earth o...	Peter B. de Selding	2009-09-21	satellite earth o...
japans htv1 cargo...	https://spacenews...	Japan's first H-2...	SpaceNews Staff	2009-09-21	japans first h2 t...
spacex says drago...	https://spacenews...	WASHINGTON — Spac...	Amy Klamper	2009-09-24	washington space...
minotaur 4 concer...	https://spacenews...	WASHINGTON — A pl...	Amy Klamper	2009-10-06	washington a pla...

小組分工

學號、姓名	貢獻比例	工作內容
楊明哲	100%	程式編寫、除錯以及文書處理