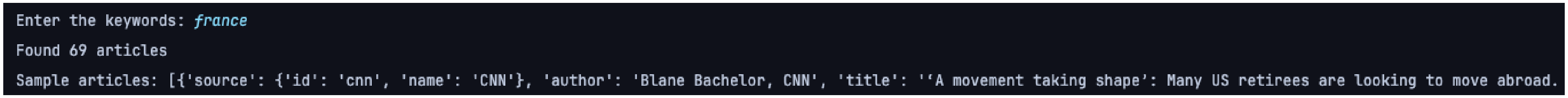
Midterm Project

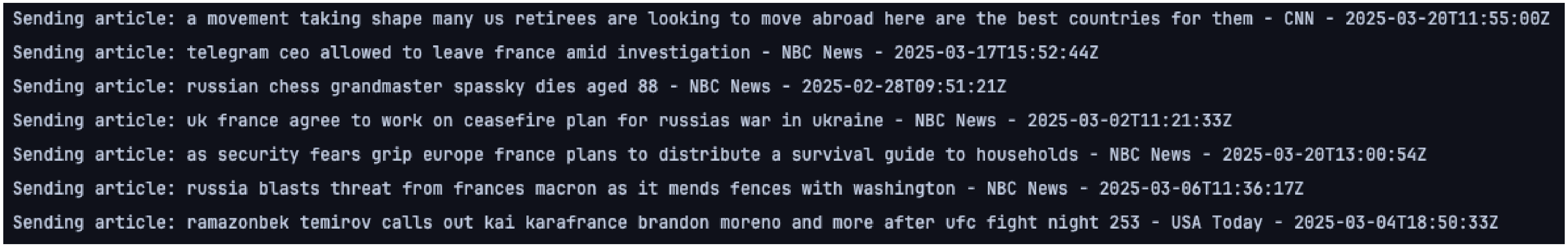
Getting the news from News API



```
Enter the keywords: france
Found 69 articles
Sample articles: [{'source': {'id': 'cnn', 'name': 'CNN'}, 'author': 'Blane Bachelor, CNN', 'title': ''A movement taking shape': Many US retirees are looking to move abroad.
```

For this pipeline, we received the articles found for the inserted keywords. To get the news, we use the unofficial library of News API to fetch any news related to the keyword and parameters configured. In this case, using the keyword "France" we obtained 69 results.

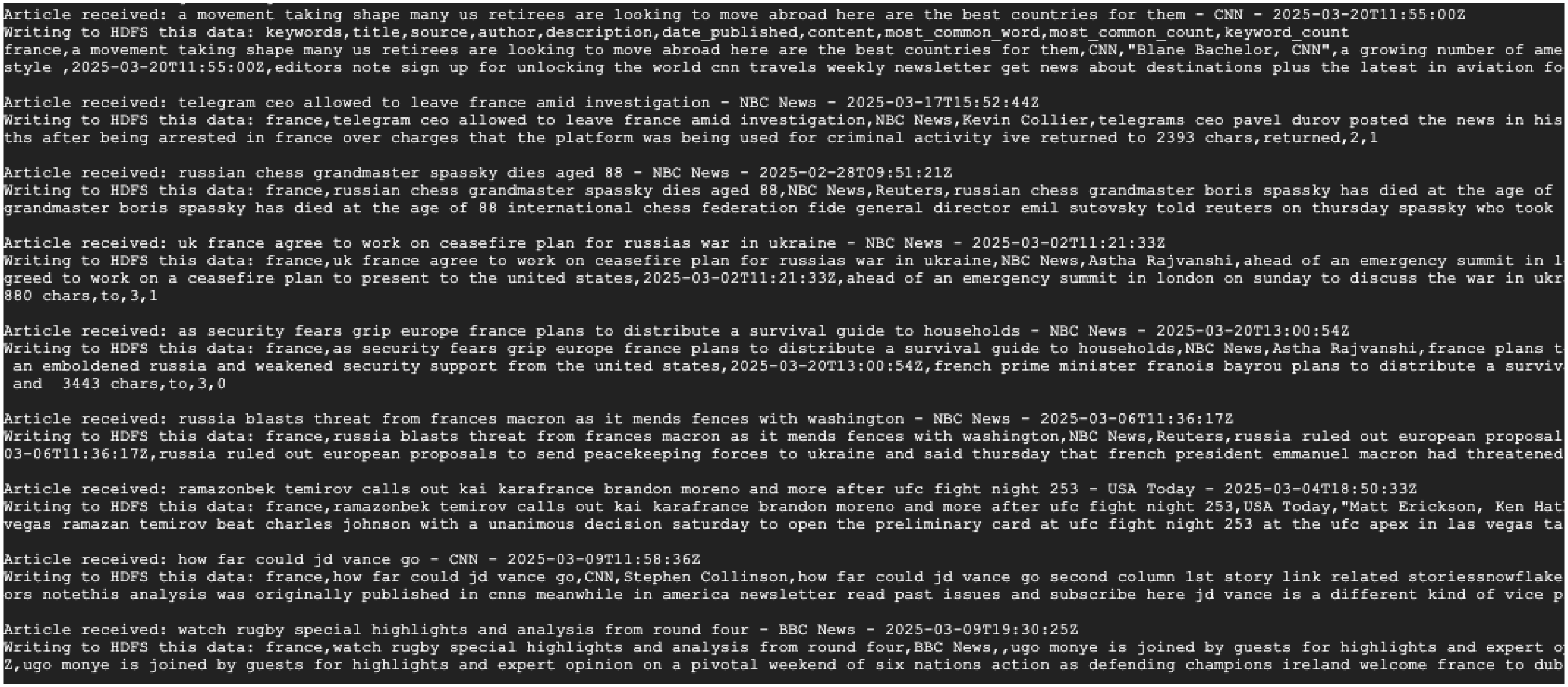
Sending news through the producer



```
Sending article: a movement taking shape many us retirees are looking to move abroad here are the best countries for them - CNN - 2025-03-20T11:55:00Z
Sending article: telegram ceo allowed to leave france amid investigation - NBC News - 2025-03-17T15:52:44Z
Sending article: russian chess grandmaster spassky dies aged 88 - NBC News - 2025-02-28T09:51:21Z
Sending article: uk france agree to work on ceasefire plan for russias war in ukraine - NBC News - 2025-03-02T11:21:33Z
Sending article: as security fears grip europe france plans to distribute a survival guide to households - NBC News - 2025-03-20T13:00:54Z
Sending article: russia blasts threat from frances macron as it mends fences with washington - NBC News - 2025-03-06T11:36:17Z
Sending article: ramazonbek temirov calls out kai karafrance brandon moreno and more after ufc fight night 253 - USA Today - 2025-03-04T18:50:33Z
```

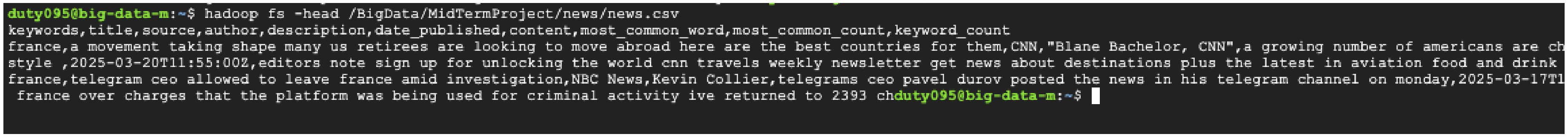
After retrieving the news from the API, we send it through the producer. Before sending it, to send the less useless data without losing significance, we perform some transformations in the title, content, and description to remove all non-alphanumeric characters, making the data sent less heavy. After performing that text processing operation, we proceed to send the article through the pipeline.

Receiving the news to the consumer and saving it into HDFS



```
Article received: a movement taking shape many us retirees are looking to move abroad here are the best countries for them - CNN - 2025-03-20T11:55:00Z
Writing to HDFS this data: keywords,title,source,author,description,date_published,content,most_common_word,most_common_count,keyword_count
france,a movement taking shape many us retirees are looking to move abroad here are the best countries for them,CNN,"Blane Bachelor, CNN",a growing number of ame
style ,2025-03-20T11:55:00Z,editors note sign up for unlocking the world cnn travels weekly newsletter get news about destinations plus the latest in aviation fo

Article received: telegram ceo allowed to leave france amid investigation - NBC News - 2025-03-17T15:52:44Z
Writing to HDFS this data: france,telegram ceo allowed to leave france amid investigation,NBC News,Kevin Collier,telegrams ceo pavel durov posted the news in his
ths after being arrested in france over charges that the platform was being used for criminal activity ive returned to 2393 chars,returned,2,1

Article received: russian chess grandmaster spassky dies aged 88 - NBC News - 2025-02-28T09:51:21Z
Writing to HDFS this data: france,russian chess grandmaster spassky dies aged 88,NBC News,Reuters,russian chess grandmaster boris spassky has died at the age of
grandmaster boris spassky has died at the age of 88 international chess federation fide general director emil sutovsky told reuters on thursday spassky who took

Article received: uk france agree to work on ceasefire plan for russias war in ukraine - NBC News - 2025-03-02T11:21:33Z
Writing to HDFS this data: france,uk france agree to work on ceasefire plan for russias war in ukraine,NBC News,Astha Rajvanshi,ahead of an emergency summit in l
greed to work on a ceasefire plan to present to the united states,2025-03-02T11:21:33Z,ahead of an emergency summit in london on sunday to discuss the war in ukr
880 chars,to,3,1

Article received: as security fears grip europe france plans to distribute a survival guide to households - NBC News - 2025-03-20T13:00:54Z
Writing to HDFS this data: france,as security fears grip europe france plans to distribute a survival guide to households,NBC News,Astha Rajvanshi,france plans t
 an emboldened russia and weakened security support from the united states,2025-03-20T13:00:54Z,french prime minister franois bayrou plans to distribute a surviv
 and  3443 chars,to,3,0

Article received: russia blasts threat from frances macron as it mends fences with washington - NBC News - 2025-03-06T11:36:17Z
Writing to HDFS this data: france,russia blasts threat from frances macron as it mends fences with washington,NBC News,Reuters,russia ruled out european proposal
03-06T11:36:17Z,russia ruled out european proposals to send peacekeeping forces to ukraine and said thursday that french president emmanuel macron had threatened

Article received: ramazonbek temirov calls out kai karafrance brandon moreno and more after ufc fight night 253 - USA Today - 2025-03-04T18:50:33Z
Writing to HDFS this data: france,ramazonbek temirov calls out kai karafrance brandon moreno and more after ufc fight night 253,USA Today,"Matt Erickson, Ken Hat
vegas ramazan temirov beat charles johnson with a unanimous decision saturday to open the preliminary card at ufc fight night 253 at the ufc apex in las vegas ta

Article received: how far could jd vance go - CNN - 2025-03-09T11:58:36Z
Writing to HDFS this data: france,how far could jd vance go,CNN,Stephen Collinson,how far could jd vance go second column 1st story link related storiessnowflake
ors notethis analysis was originally published in cnns meanwhile in america newsletter read past issues and subscribe here jd vance is a different kind of vice p

Article received: watch rugby special highlights and analysis from round four - BBC News - 2025-03-09T19:30:25Z
Writing to HDFS this data: france,watch rugby special highlights and analysis from round four,BBC News,,ugo monye is joined by guests for highlights and expert o
Z,ugo monye is joined by guests for highlights and expert opinion on a pivotal weekend of six nations action as defending champions ireland welcome france to dub
```

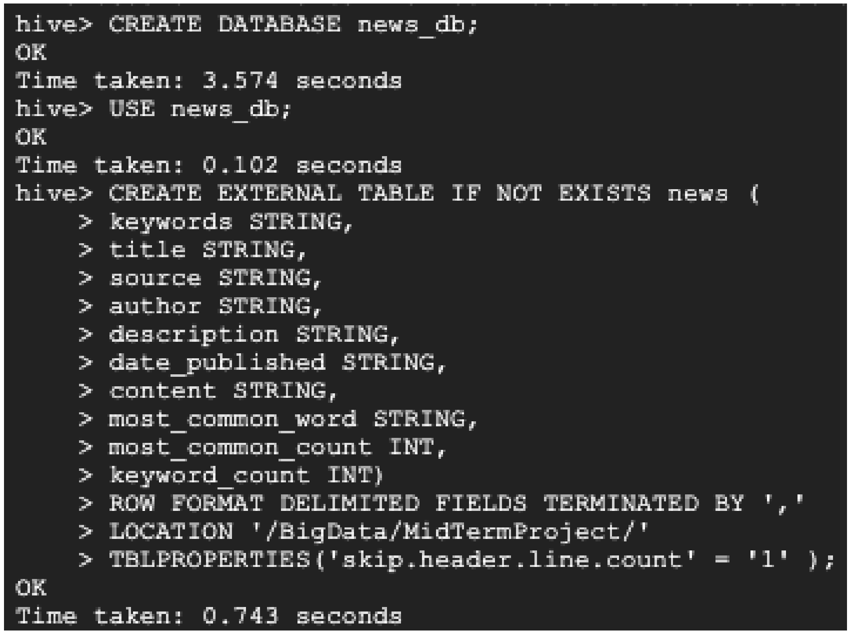
We can check here that the consumer is receiving the news sent by the producer. After receiving the information, the consumer performs some operations to extract insightful observations from the article, including the most common word, its number of occurrences, and the number of occurrences of the keyword. After processing, the data is saved in an HDFS file.



```
duty095@big-data-m:~$ hadoop fs -head /BigData/MidTermProject/news/news.csv
keywords,title,source,author,description,date_published,content,most_common_word,most_common_count,keyword_count
france,a movement taking shape many us retirees are looking to move abroad here are the best countries for them,CNN,"Blane Bachelor, CNN",a growing number of americans are ch
style ,2025-03-20T11:55:00Z,editors note sign up for unlocking the world cnn travels weekly newsletter get news about destinations plus the latest in aviation food and drink
france,telegram ceo allowed to leave france amid investigation,NBC News,Kevin Collier,telegrams ceo pavel durov posted the news in his telegram channel on monday,2025-03-17T1
 france over charges that the platform was being used for criminal activity ive returned to 2393 chduty095@big-data-m:~$
```

The image shows that the data is stored in a CSV file (news.csv). The image also shows part of the data stored in the CSV file. This data will be loaded into the hive table to gain some insights.

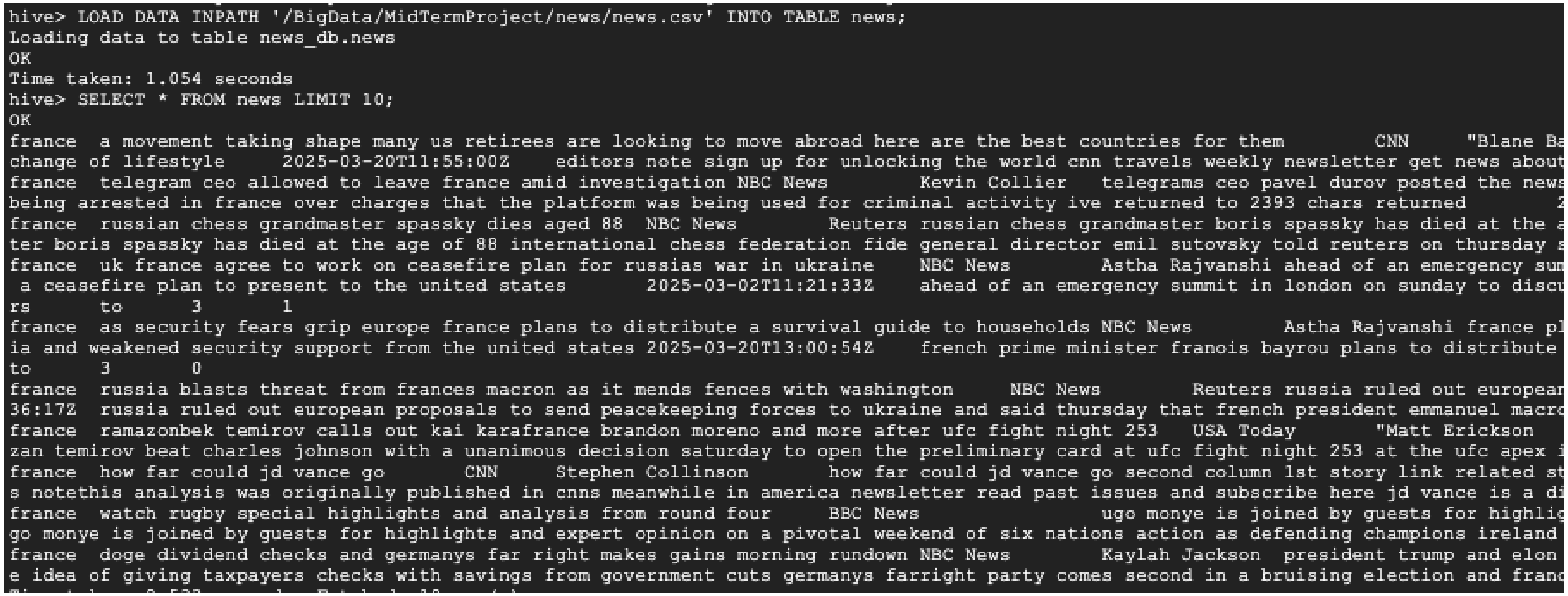
Setting up the Hive database and table



```
hive> CREATE DATABASE news_db;
OK
Time taken: 3.574 seconds
hive> USE news_db;
OK
Time taken: 0.102 seconds
hive> CREATE EXTERNAL TABLE IF NOT EXISTS news (
    > keywords STRING,
    > title STRING,
    > source STRING,
    > author STRING,
    > description STRING,
    > date_published STRING,
    > content STRING,
    > most_common_word STRING,
    > most_common_count INT,
    > keyword_count INT)
    > ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
    > LOCATION '/BigData/MidTermProject/'
    > TBLPROPERTIES('skip.header.line.count' = '1' );
OK
Time taken: 0.743 seconds
```

The schema of the table was created to load the news. This table and database store all the streamed data produced and saved by the consumer.

Loading the data generated

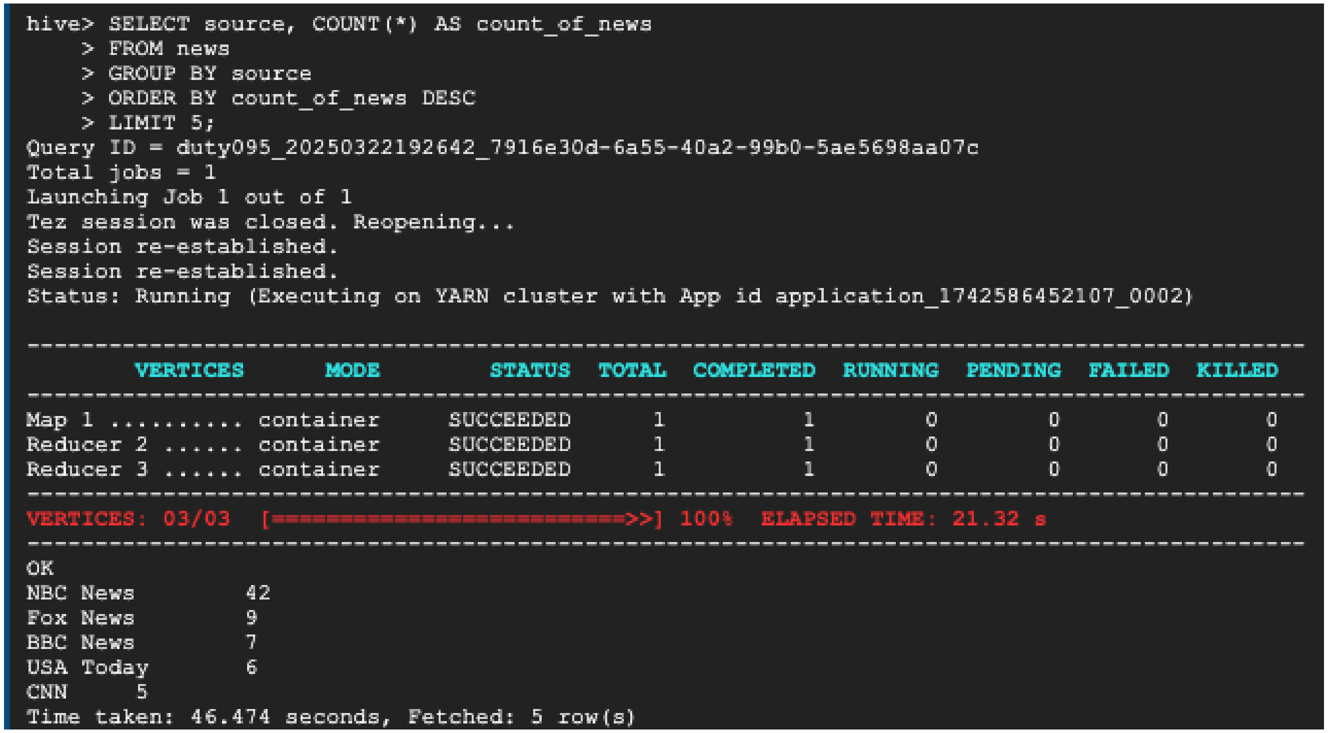


```
hive> LOAD DATA INPATH '/BigData/MidTermProject/news/news.csv' INTO TABLE news;
Loading data to table news_db.news
OK
Time taken: 1.054 seconds
hive> SELECT * FROM news LIMIT 10;
OK
france  a movement taking shape many us retirees are looking to move abroad here are the best countries for them        CNN     "Blane Ba
change of lifestyle     2025-03-20T11:55:00Z    editors note sign up for unlocking the world cnn travels weekly newsletter get news about
france  telegram ceo allowed to leave france amid investigation NBC News        Kevin Collier   telegrams ceo pavel durov posted the news
being arrested in france over charges that the platform was being used for criminal activity ive returned to 2393 chars returned        2
france  russian chess grandmaster spassky dies aged 88  NBC News        Reuters russian chess grandmaster boris spassky has died at the a
ter boris spassky has died at the age of 88 international chess federation fide general director emil sutovsky told reuters on thursday s
france  uk france agree to work on ceasefire plan for russias war in ukraine    NBC News        Astha Rajvanshi ahead of an emergency sum
 a ceasefire plan to present to the united states       2025-03-02T11:21:33Z    ahead of an emergency summit in london on sunday to discu
rs      to      3       1
france  as security fears grip europe france plans to distribute a survival guide to households NBC News        Astha Rajvanshi france pl
ia and weakened security support from the united states 2025-03-20T13:00:54Z    french prime minister franois bayrou plans to distribute
to      3       0
france  russia blasts threat from frances macron as it mends fences with washington     NBC News        Reuters russia ruled out european
36:17Z  russia ruled out european proposals to send peacekeeping forces to ukraine and said thursday that french president emmanuel macro
france  ramazonbek temirov calls out kai karafrance brandon moreno and more after ufc fight night 253   USA Today       "Matt Erickson
zan temirov beat charles johnson with a unanimous decision saturday to open the preliminary card at ufc fight night 253 at the ufc apex i
france  how far could jd vance go       CNN     Stephen Collinson       how far could jd vance go second column 1st story link related st
s notethis analysis was originally published in cnns meanwhile in america newsletter read past issues and subscribe here jd vance is a di
france  watch rugby special highlights and analysis from round four     BBC News                ugo monye is joined by guests for highlig
go monye is joined by guests for highlights and expert opinion on a pivotal weekend of six nations action as defending champions ireland
france  doge dividend checks and germanys far right makes gains morning rundown NBC News        Kaylah Jackson  president trump and elon
e idea of giving taxpayers checks with savings from government cuts germanys farright party comes second in a bruising election and franc
```

The data is loaded into the created table from the file the consumer generates to save the news. In this example, the data is already stored in the table. With this data, insights can be generated to check for useful information.

Insights

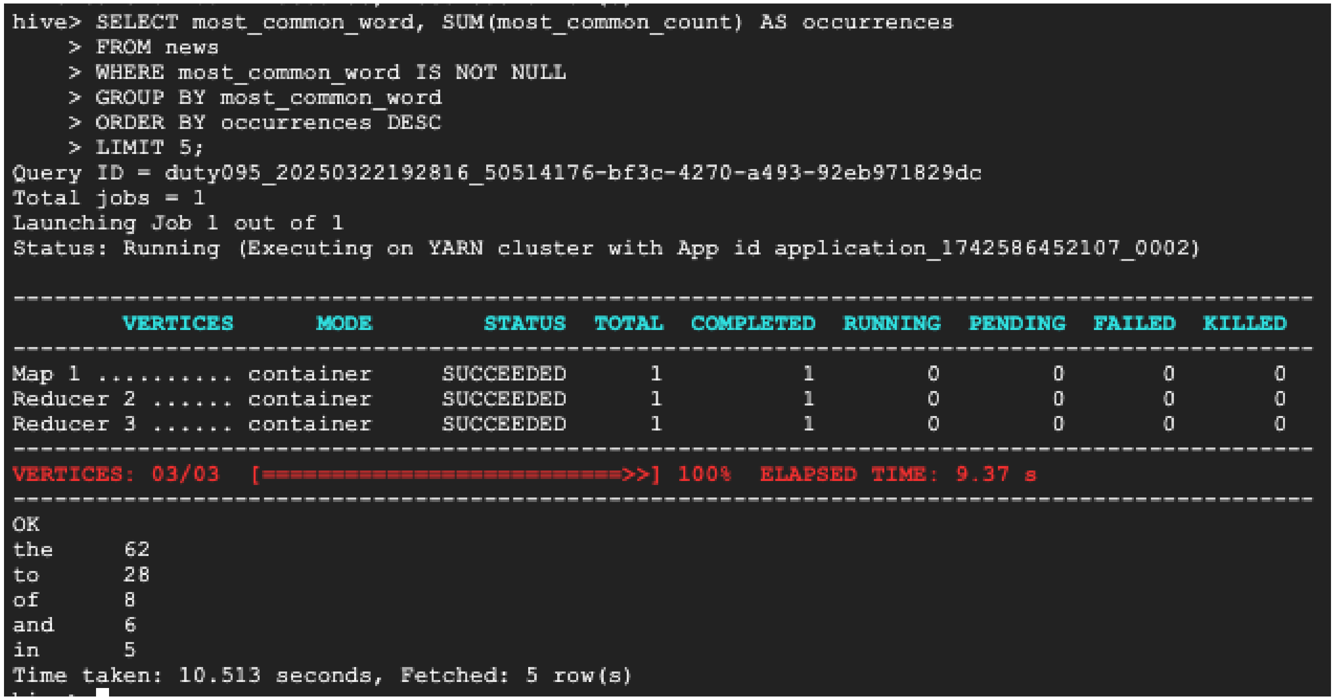
Top 5 sources by news count



```
hive> SELECT source, COUNT(*) AS count_of_news
    > FROM news
    > GROUP BY source
    > ORDER BY count_of_news DESC
    > LIMIT 5;
Query ID = duty095_20250322192642_7916e30d-6a55-40a2-99b0-5ae5698aa07c
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1742586452107_0002)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 21.32 s
----------------------------------------------------------------------------------------------
OK
NBC News        42
Fox News        9
BBC News        7
USA Today       6
CNN     5
Time taken: 46.474 seconds, Fetched: 5 row(s)
```

The first insight helps us understand which sources or media outlets contribute more to the information. For example, in this insight, NBC News is by far the one that contributes the most to this data.

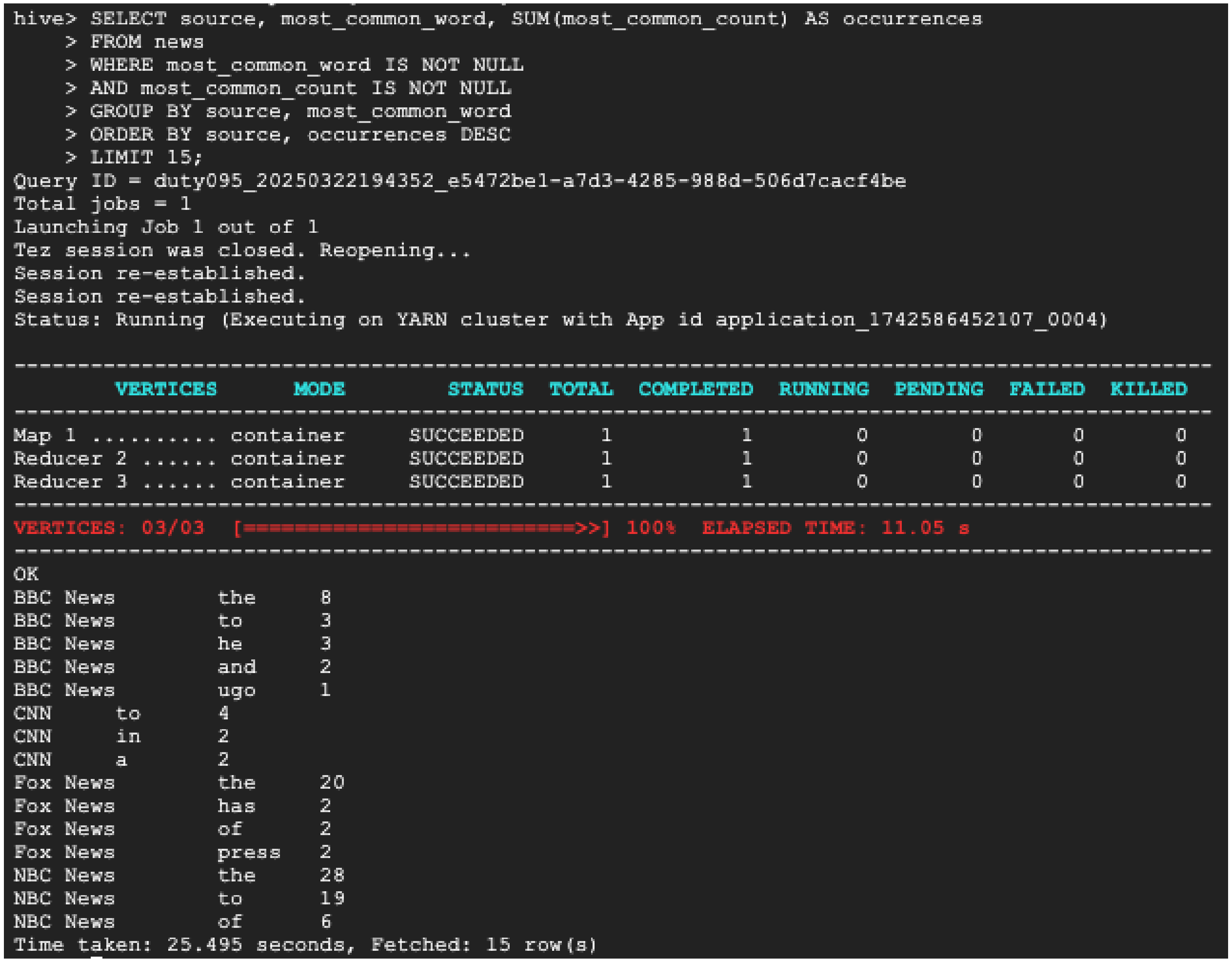
Top 5 used words in the news



```
hive> SELECT most_common_word, SUM(most_common_count) AS occurrences
    > FROM news
    > WHERE most_common_word IS NOT NULL
    > GROUP BY most_common_word
    > ORDER BY occurrences DESC
    > LIMIT 5;
Query ID = duty095_20250322192816_50514176-bf3c-4270-a493-92eb971829dc
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1742586452107_0002)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 9.37 s
----------------------------------------------------------------------------------------------
OK
the     62
to      28
of      8
and     6
in      5
Time taken: 10.513 seconds, Fetched: 5 row(s)
```

The second insight reveals the most used word in media, revealing potential biases to one ideology. In this example, the most common words found in the articles are common words like prepositions.

Top 15 used words per source



```
hive> SELECT source, most_common_word, SUM(most_common_count) AS occurrences
    > FROM news
    > WHERE most_common_word IS NOT NULL
    > AND most_common_count IS NOT NULL
    > GROUP BY source, most_common_word
    > ORDER BY source, occurrences DESC
    > LIMIT 15;
Query ID = duty095_20250322194352_e5472be1-a7d3-4285-988d-506d7cacf4be
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1742586452107_0004)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 11.05 s
----------------------------------------------------------------------------------------------
OK
BBC News        the     8
BBC News        to      3
BBC News        he      3
BBC News        and     2
BBC News        ugo     1
CNN     to      4
CNN     in      2
CNN     a       2
Fox News        the     20
Fox News        has     2
Fox News        of      2
Fox News        press   2
NBC News        the     28
NBC News        to      19
NBC News        of      6
Time taken: 25.495 seconds, Fetched: 15 row(s)
```

A more advanced insight of the second insight can get the most frequent words by source. For example, BBC and Fox News use the pronoun “the” as the most frequent word in their news.