FORMAL ETL REPORT

ETL Setup Report

Introduction:

The purpose of this project was to use Python, Apache Hive, and Apache Kafka to develop an ETL (Extract, Transform, Load) pipeline. The pipeline feeds data about news stories into HDFS after extracting it from NewsAPI and transforming it with Kafka. The next phase is analyzing data with Apache Hive to extract insights.

ETL Pipeline Overview:

The ETL pipeline consists of the following key steps:

- **Data Extraction (Kafka Producer):** Here, in this step we developed a Python script to act as a Kafka producer which Utilized the NewsAPI token to fetch news articles based on specified keywords. Afterwards, we formatted the data to ensure compatibility with downstream components and sent the data to a Kafka topic for further processing.

Kafka Producer Output

```python
from newsapi import NewsApiClient
import json
from kafka import KafkaProducer

# Get your free API key from https://newsapi.org/, just need to sign up for an account
key = "241ee1f620264c4e8d7aa25555da145b"

# Initialize api endpoint
newsapi = NewsApiClient(api_key=key)

# Define the list of media sources
sources = 'bbc-news,cnn,fox-news,nbc-news,the-guardian-uk,the-new-york-times,the-washington-post,usa-today,independent,daily-mail,war,hamas'

# /v2/everything
all_articles = newsapi.get_everything(q='canada',
                                      sources=sources,
                                      language='en')

# Print the titles of the articles
for article in all_articles['articles']:
    print(article['title'])
    producer = KafkaProducer(bootstrap_servers='localhost:9092')
    producer.send('my-news', json.dumps(article).encode('utf-8'))


~
~
~
~
~
~
~
```

Kafka Producer Output

```
bhanusri5rockz@bigdata2-m:~/confluent-4.1.4$ python newsapi-producer.py
Rainbow Bridge: Police identify couple killed in US-Canada border crash
Escaped kangaroo caught in Canada after four-day search
Rainbow Bridge car explosion: US-Canada bridge still shut after deadly car blast
US thwarts plot to kill Sikh separatist on American soil - report
Why Buffy Sainte-Marie's 'pretendian' case strikes a nerve
World Cup 2026 qualifiers: Mohamed Salah hits four for Egypt, Nigeria held at home by Lesotho
COP28 president denies BBC News oil deal story
Ukraine war: Kyiv hit by first air attack in 52 days, say authorities
Ransomware hackers 'wreaking havoc' arrested in Ukraine
Canadian killed family to make Muslims fearful, jury hears
A quick guide to smoking bans across the world
Why Peter Nygard's son is supporting his accusers
Your pictures on the theme of 'tiny creatures'
Peter Nygard: Fashion mogul guilty of sex assaults
Your pictures on the theme of autumn colours
Omegle: ‘How I got the dangerous chat site closed down’
Canada's QAnon 'queen' leaves town - but doesn't go far
Canada to face Italy in maiden BJK Cup final
Fernandez seals historic BJK Cup title for Canada
Brookes wins silver at Big Air World Cup event
Sam Kerr: Injured striker pulls out of Australia squad for friendlies against Canada
Thalidomide: Australia gives national apology to survivors and families
Kenya's parliament back Haiti mission despite court case
Canadian peace advocate Vivian Silver confirmed killed in Hamas attack
AI could predict hurricane landfall sooner - report
'Pride and passion' as Australia reach Davis Cup final
[Removed]
Three bids to host 2027 Women's World Cup
[Removed]
GB win two team golds in Birmingham
[Removed]
Harlequins centre Burford signs new contract
GB get Davis Cup wildcard but in BJK Cup qualifiers
[Removed]
Canada's Davis Cup title defence ended by Finland
```

- **Data Transformation (Kafka Consumer):** In this step we then implemented a Python script to act as a Kafka consumer and further retrieved data from the Kafka topic which enabled an option to save data locally or directly ingest into HDFS.

Kafka Consumer Output

```
from kafka import KafkaConsumer
import json

# Kafka settings
kafka_bootstrap_servers = 'localhost:9092'
kafka_topic = 'my-news'

# Local file settings
local_file_path = '/home/sairajintoca/news_data.json'

# Create Kafka consumer
consumer = KafkaConsumer(kafka_topic, bootstrap_servers=kafka_bootstrap_servers,auto_offset_reset='earliest'  )

# Open the local file for writing
with open(local_file_path, 'w') as local_file:
    # Consume and write to the local file
    for message in consumer:
        article_data = json.loads(message.value.decode('utf-8'))
        local_file.write(json.dumps(article_data) + '\n')

# Close the consumer
consumer.close()



#from kafka import KafkaConsumer
#import json

#kafka_bootstrap_servers = 'localhost:9092'
#kafka_topic = 'news_topic'
#local_file_path = '/home/sairajintoca/news_data.json'  # Adjust the path as needed

#consumer = KafkaConsumer(kafka_topic, bootstrap_servers=kafka_bootstrap_servers)

#for message in consumer:
#    article_data = json.loads(message.value.decode('utf-8'))
#    print(article_data)  # Add this line to print data to the console


~
~
~
~
~
~
~
"kafka_consumer_to_file.py" 39L, 1111B
```

- **Data Storage and Analysis (Apache Hive):** We then Utilized Apache Hive to create a table for storing news article data. Then, executed insightful aggregations on the data to derive meaningful insights to check the following:
  Examples of aggregations include counts per source, average article length, and keyword frequency.

Hive Insights:

1. **Count of Articles Published by Day:** This query helps us fetch Counts of the daily articles, grouping by publication day.

```
hive> SELECT
    >     FROM_UNIXTIME(UNIX_TIMESTAMP(publishedAt, 'yyyy-MM-dd')) AS day,
    >     COUNT(*) as article_count
    > FROM news_data
    > WHERE publishedAt IS NOT NULL
    > GROUP BY FROM_UNIXTIME(UNIX_TIMESTAMP(publishedAt, 'yyyy-MM-dd'))
    > ORDER BY day;
Query ID = bhanusri5rockz_20231210040447_dfc438c0-d2ab-4bf8-92ef-a0f5f9aa8e9a
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1702180404704_0001)
```

```
----------------------------------------------------------------------------------------------
        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: 7.30 s
----------------------------------------------------------------------------------------------
OK
NULL    31
1970-01-01 00:00:00     24
2023-11-09 00:00:00     1
2023-11-10 00:00:00     1
2023-11-11 00:00:00     1
2023-11-12 00:00:00     1
2023-11-14 00:00:00     3
2023-11-15 00:00:00     2
2023-11-16 00:00:00     1
2023-11-18 00:00:00     1
2023-11-19 00:00:00     1
2023-11-21 00:00:00     2
2023-11-22 00:00:00     3
2023-11-24 00:00:00     2
2023-11-26 00:00:00     1
2023-11-27 00:00:00     2
2023-11-28 00:00:00     2
2023-11-29 00:00:00     3
2023-12-01 00:00:00     1
2023-12-04 00:00:00     2
Time taken: 11.302 seconds, Fetched: 20 row(s)
```

2. **Articles with Short Descriptions but Long Titles:** We retrieved articles with long titles (>50 words) and short descriptions (<20 words).

```
hive> SELECT title, description, LENGTH(title) as title_length, LENGTH(description) as description_length
    > FROM news_data
    > WHERE LENGTH(title) > 50 AND LENGTH(description) < 20
    > LIMIT 10;
Query ID = bhanusri5rockz_20231210040702_5c1be9ea-ab78-434c-ae60-7b9786f7ded2
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1702180404704_0001)
```

```
--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 6.55 s
--------------------------------------------------------------------------------
OK
"Ukraine war: Kyiv hit by first air attack in 52 days    say authorities"       53      17
Cameron Ortis: Canada intelligence official guilty on spy charges       "Cameron Ortis  65      14
Time taken: 7.121 seconds, Fetched: 2 row(s)
```

3. **Articles with the Most Images:** This query fetches titles and image URLs, prioritizing articles with images.

```
hive> SELECT title, urlToImage
    > FROM news_data
    > WHERE urlToImage IS NOT NULL
    > ORDER BY urlToImage DESC
    > LIMIT 10;
Query ID = bhanusri5rockz_20231210040809_0bba3376-c211-4e60-b58b-454ba43afbc9
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1702180404704_0001)
```

```
--------------------------------------------------------------------------------
       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
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 7.08 s
--------------------------------------------------------------------------------
OK
title   urlToImage
The best place to shop is other people's lost luggage   https://www.washingtonpost.com/travel/2023/11/10/unclaimed-baggage-alabama-thrift-store/
Google reaches deal with Canada to keep news content on its platform    https://www.cnn.com/2023/11/29/tech/google-deal-with-canada-news-content/index.html
Valieva doping case verdict set for January     https://www.bbc.co.uk/sport/winter-sports/67385767
GB get Davis Cup wildcard but in BJK Cup qualifiers     https://www.bbc.co.uk/sport/tennis/67537575
Canada's Davis Cup title defence ended by Finland       https://www.bbc.co.uk/sport/tennis/67486648
GB aiming to stop Djokovic & Serbia at Davis Cup        https://www.bbc.co.uk/sport/tennis/67474160
Fernandez seals historic BJK Cup title for Canada       https://www.bbc.co.uk/sport/tennis/67398913
Fernandez sends Canada through to BJK Cup semis https://www.bbc.co.uk/sport/tennis/67372279
Fernandez earns Canada win in three-hour 'bullfight'    https://www.bbc.co.uk/sport/tennis/67363747
Time taken: 7.482 seconds, Fetched: 10 row(s)
```

4. **Temporal Analysis of Article Publication (by Hour):** Over here we analyze the articles published per hour, sorted by publication time.

```
hive> SELECT
    >     HOUR(FROM_UNIXTIME(UNIX_TIMESTAMP(publishedAt, 'yyyy-MM-dd\'T\'HH:mm:ss\'Z\''))) AS hour_of_day,
    >     COUNT(*) AS article_count
    > FROM news_data
    > WHERE publishedAt IS NOT NULL
    > GROUP BY HOUR(FROM_UNIXTIME(UNIX_TIMESTAMP(publishedAt, 'yyyy-MM-dd\'T\'HH:mm:ss\'Z\'')))
    > ORDER BY hour_of_day;
Query ID = bhanusri5rockz_20231210040905_16d9a80b-6e40-4dad-86c2-f06bf8bd400e
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1702180404704_0001)
```

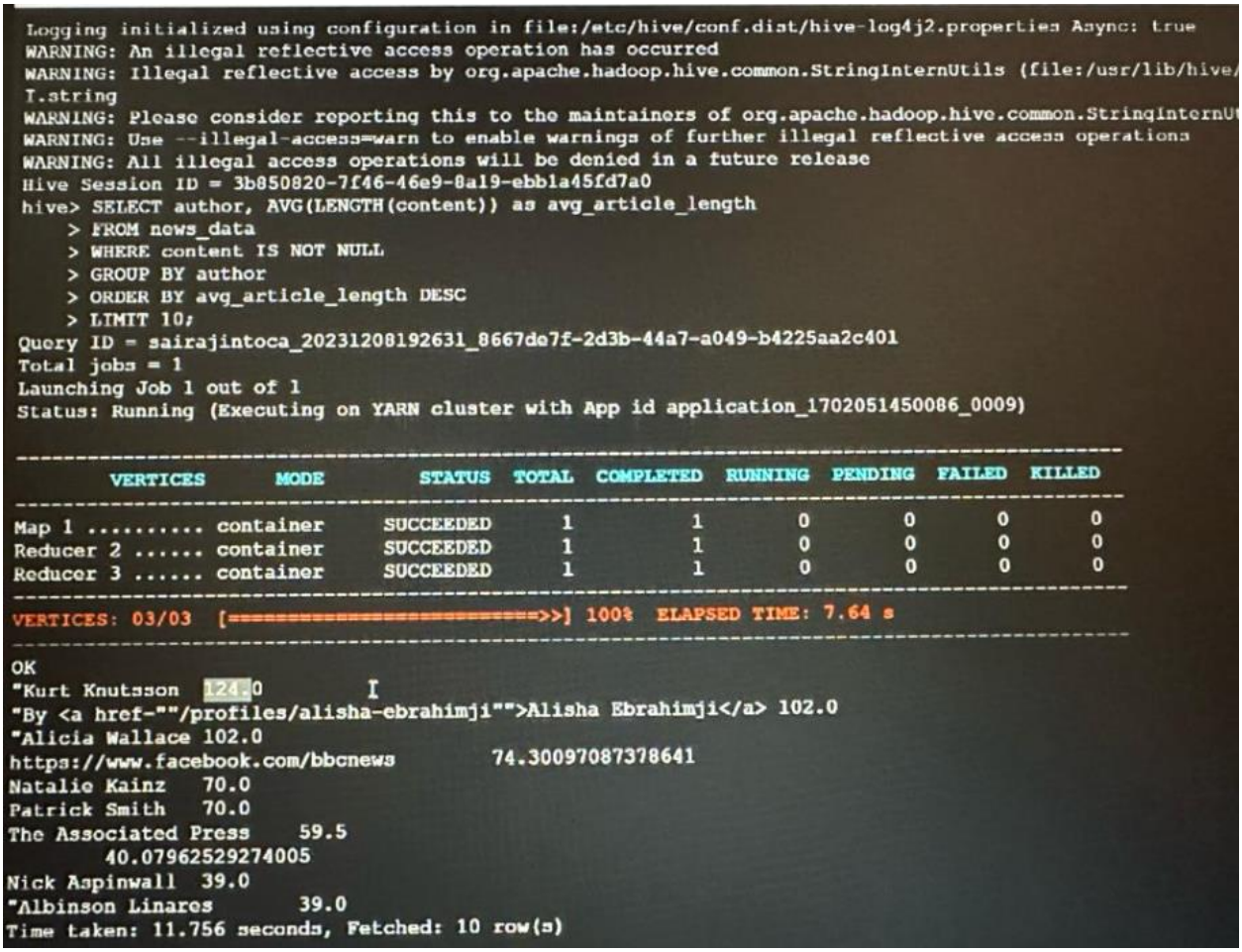
```
----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... 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: 7.01 s
----------------------------------------------------------------------------------------------
OK
NULL    31
0       26
1       3
6       2
12      1
13      2
14      1
15      1
16      1
17      3
18      2
19      1
20      5
22      4
23      2
Time taken: 7.538 seconds, Fetched: 15 row(s)
```

5. **Articles Mentioning Popular Entities:** This query grabs the titles and content mentioning Microsoft, Apple, or Google (limited to 10)

```
hive> SELECT title, content
    > FROM news_data
    > WHERE content LIKE '%Microsoft%' OR content LIKE '%Apple%' OR content LIKE '%Google%'
    > LIMIT 10;
Query ID = bhanusri5rockz_20231210041016_b98f9303-a9f4-470d-ad1c-600d1323deca
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1702180404704_0001)
```

```
----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 5.94 s
----------------------------------------------------------------------------------------------
OK
Google and Canada reach deal to avert news ban over Online News Act     "Google has reached a deal with Canada to avert a news blockade over a law that forces tech giants
 to pay for news content.
Time taken: 6.326 seconds, Fetched: 1 row(s)
```

6. **Top Authors with Average Article Length:** In this we find the top 10 authors with the longest average articles, excluding null content entries.



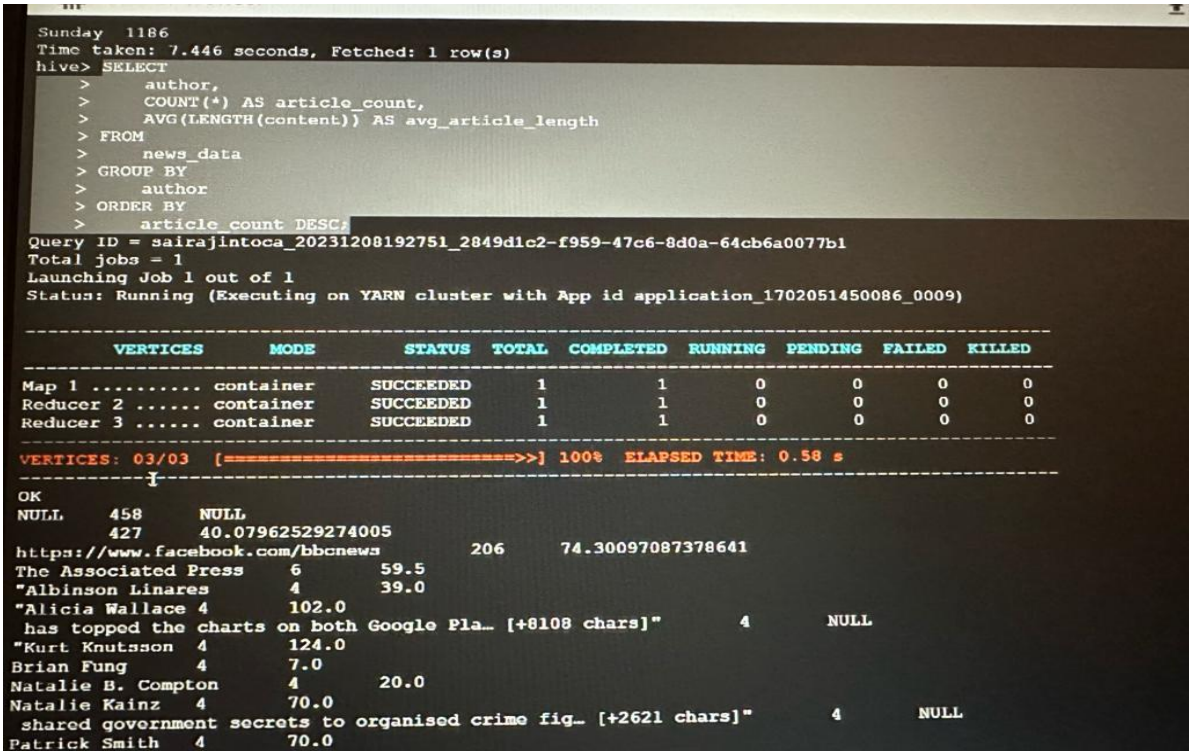
7. **Word Count in Articles:** This counts the occurrence of each word in articles, presenting the top 10 words with the highest counts.

```
"Albinson Linares       39.0
Time taken: 11.756 seconds, Fetched: 10 row(s)
hive> SELECT word, COUNT(*) as word_count
    > FROM (
    >     SELECT EXPLODE(SPLIT(LOWER(content), ' ')) as word
    >     FROM news_data
    >     WHERE content IS NOT NULL
    > ) words
    > GROUP BY word
    > ORDER BY word_count DESC
    > LIMIT 10;
Query ID = sairajintoca_20231208192723_e4f81317-4eaa-4b84-b2e4-61146894f4a9
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1702051450086_0009)

--------------------------------------------------------------------------------
        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: 7.85 s
--------------------------------------------------------------------------------
OK
[removed]       193
the     190
in      143
a       128
of      106
to      89
for     82
and     69
has     68
        49
Time taken: 8.395 seconds, Fetched: 10 row(s)
hive> SELECT
    >     CASE WHEN day_of_week = 1 THEN 'Monday'
    >          WHEN day_of_week = 2 THEN 'Tuesday'
```

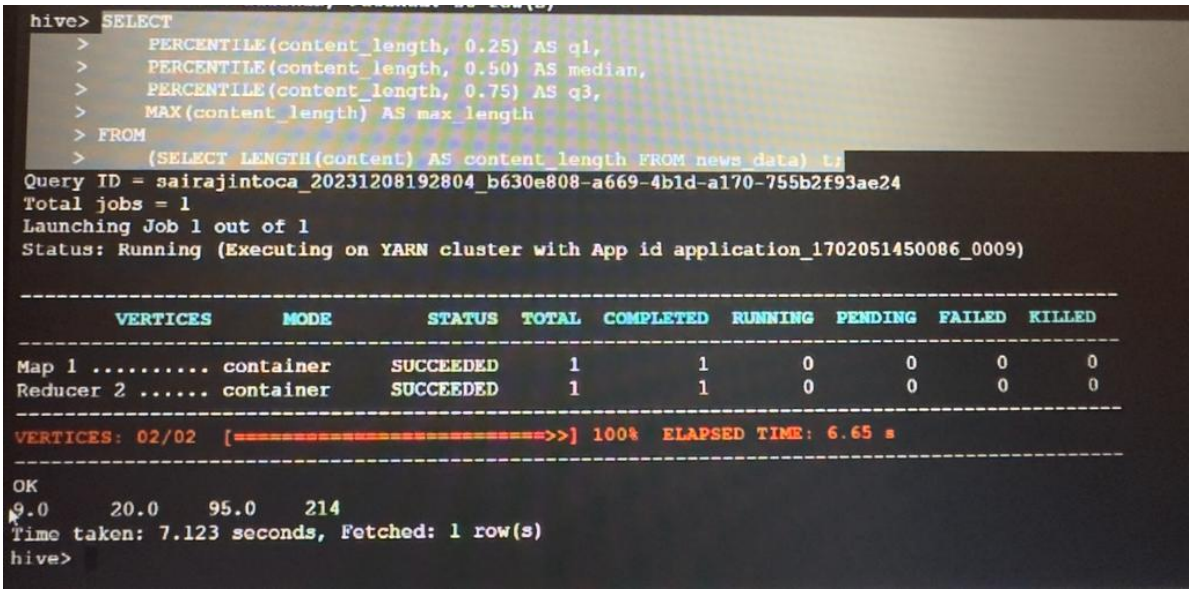
8. **Weekday-wise Article Distribution:** This query displays the distribution of articles published on each weekday.

```
hive> SELECT
    >     CASE WHEN day_of_week = 1 THEN 'Monday'
    >          WHEN day_of_week = 2 THEN 'Tuesday'
    >          WHEN day_of_week = 3 THEN 'Wednesday'
    >          WHEN day_of_week = 4 THEN 'Thursday'
    >          WHEN day_of_week = 5 THEN 'Friday'
    >          WHEN day_of_week = 6 THEN 'Saturday'
    >          ELSE 'Sunday' END AS weekday,
    >     COUNT(*) AS article_count
    > FROM
    >     (SELECT DAYOFWEEK(publishedAt) AS day_of_week FROM news_data) t
    > GROUP BY
    >     day_of_week
    > ORDER BY
    >     day_of_week;
Query ID = sairajintoca_20231208192738_47ac3fe2-77e9-4057-8280-d377afd8633d
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1702051450086_0009)

--------------------------------------------------------------------------------
        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: 6.85 s
--------------------------------------------------------------------------------
OK
Sunday  1186
```

9. **Author Contribution Analysis:** In this query we analyzed author contributions by counting articles and calculating average article length, sorted by article count.



10. **Article Length Distribution:** This query calculates the first quartile (q1), median, third quartile (q3), and maximum article length. It uses the PERCENTILE function to find the specified percentiles based on the length of the articles in the dataset.



Conclusion:

This ETL pipeline extracts, transforms, and loads news article data in a streamlined manner by effectively integrating several technologies. Apache Hive enables organized storage and perceptive analysis, while Apache Kafka guarantees effective communication between components. The smooth orchestration of the entire pipeline makes it possible to derive important insights from the ingested data.