Data Collection and Preparation - Final Project Report

Team Members: Alibekov Bakytzhan, Sharipov Damir, Seilkhan Aidariya

1. API Justification

Chosen API: We have been using api from newsapi.api and it looks like eventregistry.com This API was selected because: (1) **Frequent updates** - provides real-time news articles with continuous publishing throughout the day, our implementation fetches technology articles every 1 minute; (2) **Stable and documented** - well-established platform with comprehensive documentation; (3) **Structured JSON** - returns well-structured responses with article metadata, sentiment scores, and relevance metrics; (4) **Real data** - genuine news articles from real sources with valuable attributes for analytics;

2. Kafka Topic Schema

Topic Name: `raw_events`
Each message contains a timestamp and raw API response data. Example structure:

Field	Type	Description
timestamp	string	ISO 8601 timestamp when data was fetched
data	object	Raw JSON response from EventRegistry API
data.articles.results	array	Array of article objects with metadata
results[].uri	string	Unique article identifier (e.g., "9008770108")
results[].lang	string	Article language code (e.g., "eng", "spa", "fra")
results[].date	string	Publication date (YYYY-MM-DD)
results[].time	string	Publication time (HH:MM:SS)
results[].dateTime	string	Publication datetime (ISO 8601)
results[].dataType	string	Type of data (e.g., "news")
results[].url	string	Original article URL
results[].title	string	Article headline/title
results[].body	string	Full article text content
results[].source	object	Source information with uri and title
results[].source.uri	string	Source identifier (e.g., "eff.org")
results[].source.title	string	Source name (e.g., "Electronic Frontier Foundation")
results[].authors	array	Array of author objects
results[].image	string	Article image URL (if available)
results[].sentiment	number	Sentiment score in range [-1.0, 1.0]
results[].wgt	number	Article weight/importance score
results[].relevance	number	Topic relevance score

3. Data Cleaning Rules

All operations use **Pandas vectorized operations** (no loops): (1) Extract nested fields with df.apply(); (2) Rename columns with df.rename(); (3) Remove nulls with df.dropna(); (4) Remove duplicates with df.drop_duplicates(); (5) Fill missing values with df.fillna(); (6) Clean text with str.strip(); (7) Convert types with pd.to_numeric(); (8) Normalize sentiment with df.clip(-1.0, 1.0); (9) Add timestamp with pd.Timestamp.now().

4. SQLite Database Schema

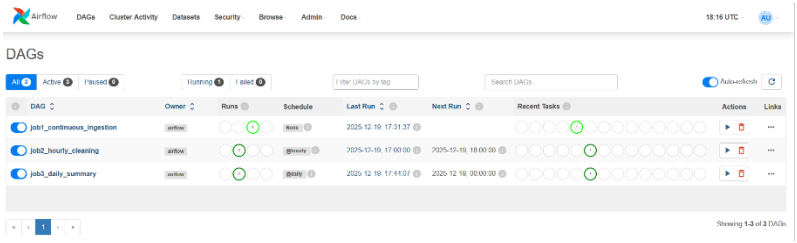
Table 1: events (Cleaned Data)

Column Name	Data Type	Constraints	Description
article_uri	TEXT	PRIMARY KEY	Unique article identifier
lang	TEXT		Article language code (eng, spa, fra, etc.)
datetime	TEXT		Publication date and time (ISO 8601 format)
data_type	TEXT		Type of data (e.g., "news")
url	TEXT		Original article URL
title	TEXT		Article headline/title
body	TEXT		Full article text content
source_uri	TEXT		Source identifier (e.g., "bbc.co.uk")
image_url	TEXT		Article image URL (if available)
sentiment	REAL		Sentiment score in range [-1.0, 1.0]
wgt	REAL		Article weight/importance score
relevance	REAL		Topic relevance score
created_at	TIMESTAMP	DEFAULT CURRENT_TIMESTAMP	Record Insertion timestamp

Table 2: daily_summary (Analytics)

Column Name	Data Type	Constraints	Description
id	INTEGER	PRIMARY KEY AUTOINCREMENT	Auto-increment ID
summary_date	TEXT	UNIQUE	Date of summary (YYYY-MM-DD)
total_articles	INTEGER		Total number of articles for the date
avg_sentiment	REAL		Average sentiment score
min_sentiment	REAL		Minimum sentiment score
max_sentiment	REAL		Maximum sentiment score
top_source	TEXT		Source URI with most articles
language_distribution	TEXT		JSON string with language counts
created_at	TIMESTAMP	DEFAULT CURRENT_TIME-STAMP	Summary creation timestamp

5. DAG Execution Logs (Screenshots)



DAG 1: Continuous Ingestion - Messages sent to Kafka every 1 minute

```
[2025-12-19, 17:31:41 UTC] {logging_mixin.py:154} INFO - Message sent to topic raw_events
[2025-12-19, 17:31:41 UTC] {logging_mixin.py:154} INFO - Successfully sent message #1
[2025-12-19, 17:32:41 UTC] {logging_mixin.py:154} INFO - Fetching data from API (message #2)...
[2025-12-19, 17:32:42 UTC] {logging_mixin.py:154} INFO - Message sent to topic raw_events
[2025-12-19, 17:32:42 UTC] {logging_mixin.py:154} INFO - Successfully sent message #2
[2025-12-19, 17:33:42 UTC] {logging_mixin.py:154} INFO - Fetching data from API (message #3)...
[2025-12-19, 17:33:43 UTC] {logging_mixin.py:154} INFO - Message sent to topic raw_events
[2025-12-19, 17:33:43 UTC] {logging_mixin.py:154} INFO - Successfully sent message #3
```

DAG 2: Hourly Cleaning and Storage - Cleaned 5 articles with Pandas

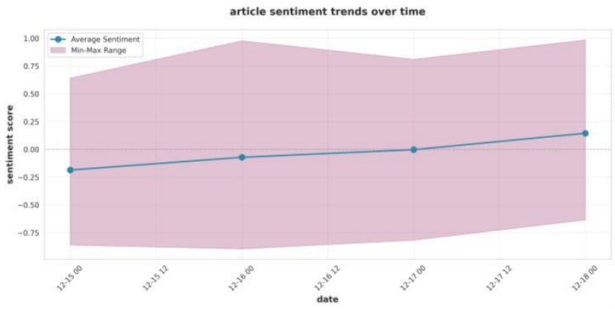
```
[2025-12-19, 17:46:58 UTC] {logging_mixin.py:154} INFO - Processing message #1
[2025-12-19, 17:46:58 UTC] {logging_mixin.py:154} INFO - Processing message #2
[2025-12-19, 17:46:58 UTC] {logging_mixin.py:154} INFO - Processing message #3
[2025-12-19, 17:47:01 UTC] {base.py:829} WARNING - Heartbeat failed for group job2_hourly_cleaning because it is rebalancing
[2025-12-19, 17:47:01 UTC] {consumer.py:348} INFO - Revoking previously assigned partitions (('raw_events', 0),) for group job2_hourly_cleaning
[2025-12-19, 17:47:01 UTC] {base.py:450} INFO - (Re-)joining group job2_hourly_cleaning
[2025-12-19, 17:47:01 UTC] {base.py:521} INFO - Elected group leader -- performing partition assignments using range
[2025-12-19, 17:47:01 UTC] {base.py:335} INFO - Successfully joined group job2_hourly_cleaning with generation 2
[2025-12-19, 17:47:01 UTC] {subscription_state.py:257} INFO - Updated partition assignment: [('raw_events', 0)]
[2025-12-19, 17:47:01 UTC] {consumer.py:245} INFO - Setting newly assigned partitions (('raw_events', 0),) for group job2_hourly_cleaning
[2025-12-19, 17:47:08 UTC] {logging_mixin.py:154} INFO - Collected 15 total articles from 3 messages
[2025-12-19, 17:47:08 UTC] {logging_mixin.py:154} INFO - Creating DataFrame...
[2025-12-19, 17:47:08 UTC] {logging_mixin.py:154} INFO - Starting data cleaning process...
[2025-12-19, 17:47:08 UTC] {logging_mixin.py:154} INFO - Initial articles count: 15
[2025-12-19, 17:47:08 UTC] {logging_mixin.py:154} INFO - Extracting source_url from nested source field...
[2025-12-19, 17:47:08 UTC] {logging_mixin.py:154} INFO - Renaming columns...
[2025-12-19, 17:47:08 UTC] {logging_mixin.py:154} INFO - Removed 5 duplicate articles
[2025-12-19, 17:47:08 UTC] {logging_mixin.py:154} INFO - Filling missing values and cleaning text fields...
[2025-12-19, 17:47:08 UTC] {logging_mixin.py:154} INFO - Removing excessive whitespace from body and title...
[2025-12-19, 17:47:08 UTC] {logging_mixin.py:154} INFO - Converting numeric fields and normalizing sentiment...
[2025-12-19, 17:47:08 UTC] {logging_mixin.py:154} INFO - Final articles count: 10
[2025-12-19, 17:47:08 UTC] {logging_mixin.py:154} INFO - Cleaned 10 articles using Pandas
[2025-12-19, 17:47:08 UTC] {logging_mixin.py:154} INFO - Saving 10 cleaned articles to database...
[2025-12-19, 17:47:08 UTC] {logging_mixin.py:154} INFO - sample of data from database:
[2025-12-19, 17:47:08 UTC] {logging_mixin.py:154} INFO - [1] Article URI: 9008770108
[2025-12-19, 17:47:08 UTC] {logging_mixin.py:154} INFO - Title: Speaking Freely: Sami Ben Gharbia
[2025-12-19, 17:47:08 UTC] {logging_mixin.py:154} INFO - Language: eng
[2025-12-19, 17:47:08 UTC] {logging_mixin.py:154} INFO - Sentiment: 0.5764705882352941
[2025-12-19, 17:47:08 UTC] {logging_mixin.py:154} INFO - DateTime: 2025-12-19T17:42:43Z
[2025-12-19, 17:47:08 UTC] {logging_mixin.py:154} INFO - [2] Article URI: 9008770445
[2025-12-19, 17:47:08 UTC] {logging_mixin.py:154} INFO - Title: 10 Information Technology Stocks Whale Activity In Today's Session - Apple (NASD...
[2025-12-19, 17:47:08 UTC] {logging_mixin.py:154} INFO - Language: eng
[2025-12-19, 17:47:08 UTC] {logging_mixin.py:154} INFO - Sentiment: 0.12156862745098044
[2025-12-19, 17:47:08 UTC] {logging_mixin.py:154} INFO - DateTime: 2025-12-19T17:42:42Z
[2025-12-19, 17:47:08 UTC] {logging_mixin.py:154} INFO - [3] Article URI: 9008769403
[2025-12-19, 17:47:08 UTC] {logging_mixin.py:154} INFO - Title: From H-1B visa to billion-dollar success: The Jyoti Bansal story
[2025-12-19, 17:47:08 UTC] {logging_mixin.py:154} INFO - Language: eng
[2025-12-19, 17:47:08 UTC] {logging_mixin.py:154} INFO - Sentiment: 0.2941176470588236
[2025-12-19, 17:47:08 UTC] {logging_mixin.py:154} INFO - DateTime: 2025-12-19T17:42:14Z
[2025-12-19, 17:47:08 UTC] {logging_mixin.py:154} INFO - Total articles in database: 60
```

DAG 3: Daily Summary - Computed analytics for 20 articles

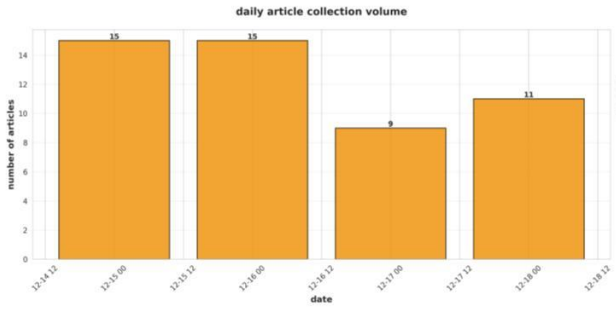
```
[2025-12-19, 17:47:51 UTC] {logging_mixin.py:154} INFO - Database tables created or verified.
[2025-12-19, 17:47:51 UTC] {logging_mixin.py:154} INFO - total evennts 60
[2025-12-19, 17:47:51 UTC] {logging_mixin.py:154} INFO - we found 5 uniques
[2025-12-19, 17:47:51 UTC] {logging_mixin.py:154} INFO -  date - 2025-12-16
[2025-12-19, 17:47:51 UTC] {logging_mixin.py:154} INFO - total - 15
[2025-12-19, 17:47:51 UTC] {logging_mixin.py:154} INFO - mean - -0.071, min - -0.895, max - 0.978
[2025-12-19, 17:47:51 UTC] {logging_mixin.py:154} INFO - top surce - techcrunch.com (7
[2025-12-19, 17:47:51 UTC] {logging_mixin.py:154} INFO - lang distribution - {"eng": 5, "fra": 6, "spa": 4}
[2025-12-19, 17:47:51 UTC] {logging_mixin.py:154} INFO - summary saved 2025-12-16
[2025-12-19, 17:47:51 UTC] {logging_mixin.py:154} INFO -  date - 2025-12-15
[2025-12-19, 17:47:51 UTC] {logging_mixin.py:154} INFO - total - 15
[2025-12-19, 17:47:51 UTC] {logging_mixin.py:154} INFO - mean - -0.186, min - -0.861, max - 0.642
[2025-12-19, 17:47:51 UTC] {logging_mixin.py:154} INFO - top surce - cnn.com (5
[2025-12-19, 17:47:51 UTC] {logging_mixin.py:154} INFO - lang distribution - {"eng": 2, "fra": 3, "spa": 10}
[2025-12-19, 17:47:51 UTC] {logging_mixin.py:154} INFO - summary saved 2025-12-15
[2025-12-19, 17:47:51 UTC] {logging_mixin.py:154} INFO -  date - 2025-12-18
[2025-12-19, 17:47:51 UTC] {logging_mixin.py:154} INFO - total - 11
[2025-12-19, 17:47:51 UTC] {logging_mixin.py:154} INFO - mean - 0.145, min - -0.634, max - 0.985
[2025-12-19, 17:47:51 UTC] {logging_mixin.py:154} INFO - top surce - reuters.com (5
[2025-12-19, 17:47:51 UTC] {logging_mixin.py:154} INFO - lang distribution - {"eng": 2, "fra": 3, "spa": 6}
[2025-12-19, 17:47:51 UTC] {logging_mixin.py:154} INFO - summary saved 2025-12-18
[2025-12-19, 17:47:51 UTC] {logging_mixin.py:154} INFO -  date - 2025-12-17
[2025-12-19, 17:47:51 UTC] {logging_mixin.py:154} INFO - total - 9
[2025-12-19, 17:47:51 UTC] {logging_mixin.py:154} INFO - mean - -0.003, min - -0.817, max - 0.811
[2025-12-19, 17:47:51 UTC] {logging_mixin.py:154} INFO - top surce - reuters.com (3
[2025-12-19, 17:47:51 UTC] {logging_mixin.py:154} INFO - lang distribution - {"eng": 4, "fra": 3, "spa": 2}
[2025-12-19, 17:47:51 UTC] {logging_mixin.py:154} INFO - summary saved 2025-12-17
[2025-12-19, 17:47:51 UTC] {logging_mixin.py:154} INFO -  date - 2025-12-19
[2025-12-19, 17:47:51 UTC] {logging_mixin.py:154} INFO - total - 10
[2025-12-19, 17:47:51 UTC] {logging_mixin.py:154} INFO - mean - 0.268, min - -0.239, max - 0.576
[2025-12-19, 17:47:51 UTC] {logging_mixin.py:154} INFO - top surce - eff.org (1
[2025-12-19, 17:47:51 UTC] {logging_mixin.py:154} INFO - lang distribution - {"eng": 10}
[2025-12-19, 17:47:51 UTC] {logging_mixin.py:154} INFO - summary saved 2025-12-19
[2025-12-19, 17:47:51 UTC] {logging_mixin.py:154} INFO -  5 date processed
[2025-12-19, 17:47:51 UTC] {logging_mixin.py:154} INFO -   date 2025-12-19
[2025-12-19, 17:47:51 UTC] {logging_mixin.py:154} INFO -   total 10
[2025-12-19, 17:47:51 UTC] {logging_mixin.py:154} INFO -   mean sentiment 0.268
[2025-12-19, 17:47:51 UTC] {logging_mixin.py:154} INFO -   top surce eff.org
[2025-12-19, 17:47:51 UTC] {logging_mixin.py:154} INFO -   langs {"eng": 10}
[2025-12-19, 17:47:51 UTC] {logging_mixin.py:154} INFO -   date 2025-12-18
[2025-12-19, 17:47:51 UTC] {logging_mixin.py:154} INFO -   total 11
[2025-12-19, 17:47:51 UTC] {logging_mixin.py:154} INFO -   mean sentiment 0.145
```

6. Data Visualizations

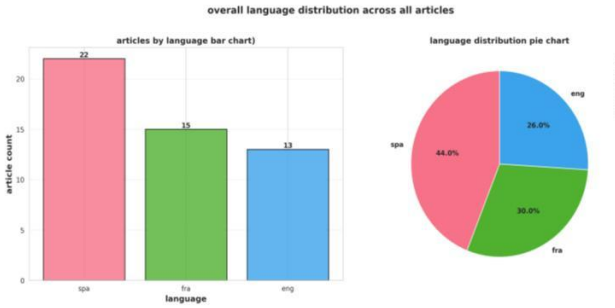
Sentiment Trends Over Time



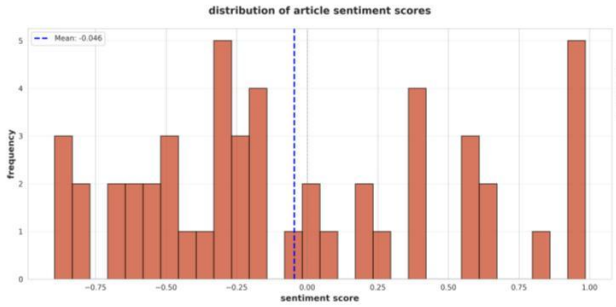
Daily Article Volume



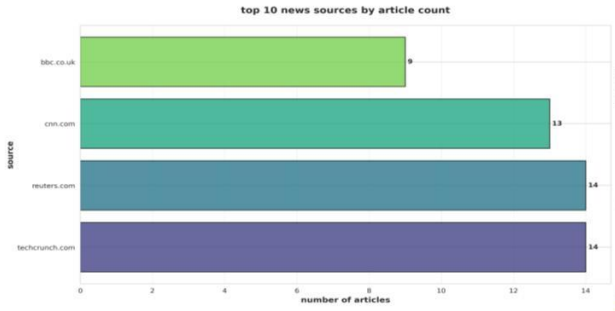
Language Distribution



Sentiment Distribution



Top 10 News Sources



Language Heatmap Across Dates