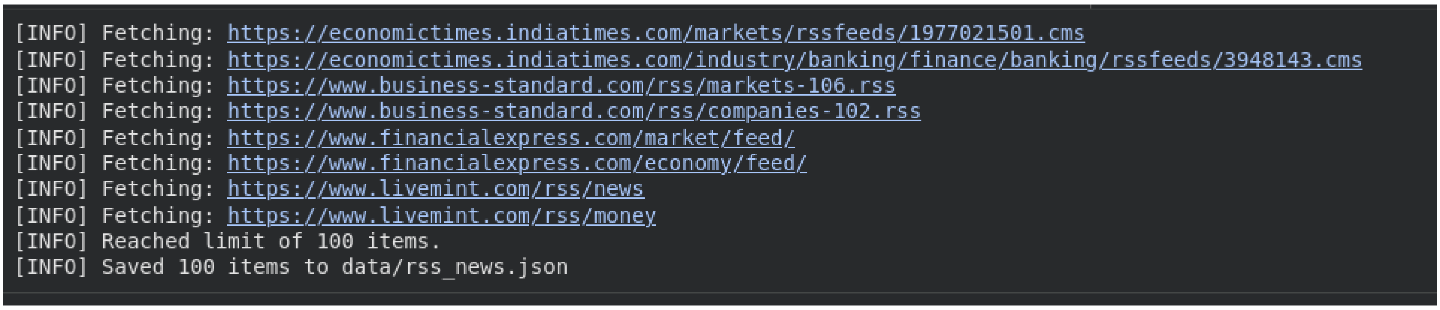
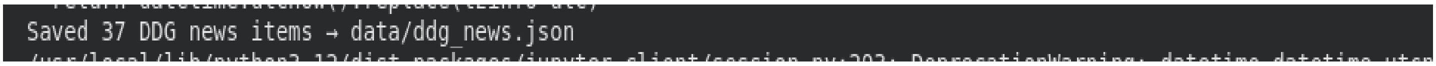
Financial News Multi-Agent Analysis System

Data Extraction for News feeds

We collect raw financial news articles, press releases, and announcements from the provided datasets. The text is cleaned, formatted, and structured so it can be processed reliably by downstream agents.



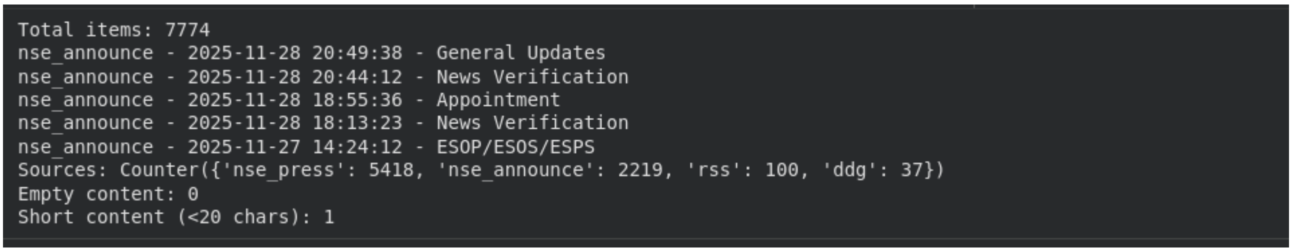
```
[INFO] Fetching: https://economictimes.indiatimes.com/markets/rssfeeds/1977021501.cms
[INFO] Fetching: https://economictimes.indiatimes.com/industry/banking/finance/banking/rssfeeds/3948143.cms
[INFO] Fetching: https://www.business-standard.com/rss/markets-106.rss
[INFO] Fetching: https://www.business-standard.com/rss/companies-102.rss
[INFO] Fetching: https://www.financialexpress.com/market/feed/
[INFO] Fetching: https://www.financialexpress.com/economy/feed/
[INFO] Fetching: https://www.livemint.com/rss/news
[INFO] Fetching: https://www.livemint.com/rss/money
[INFO] Reached limit of 100 items.
[INFO] Saved 100 items to data/rss_news.json
```



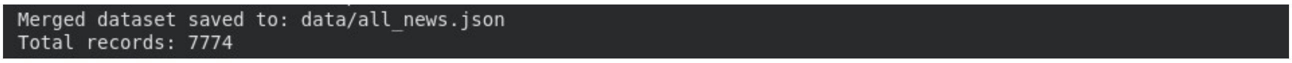
```
Saved 37 DDG news items → data/ddg_news.json
```

Data audit : understand your collection

Before running the pipeline, we analyze the dataset to understand its size, quality, duplicates, missing fields, and overall distribution. This helps validate the input data and ensures the pipeline handles it correctly.

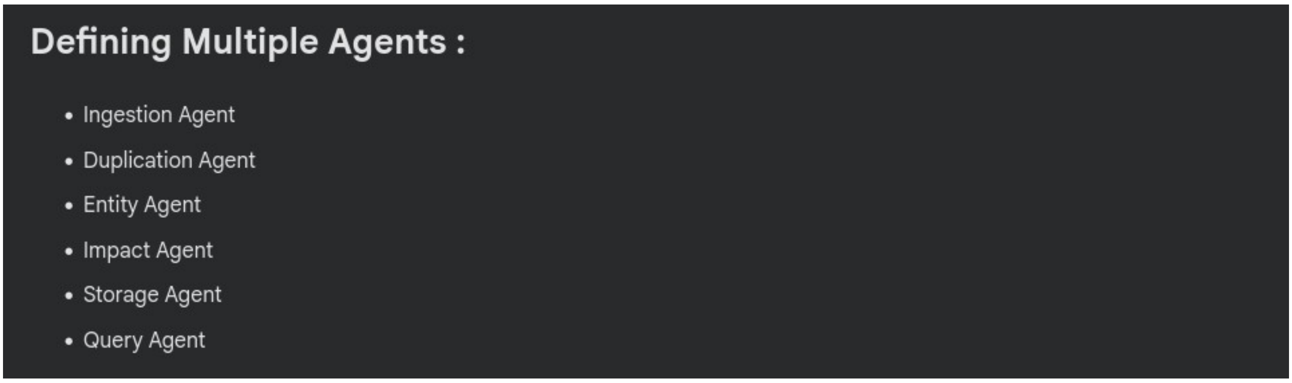


```
Total items: 7774
nse_announce - 2025-11-28 20:49:38 - General Updates
nse_announce - 2025-11-28 20:44:12 - News Verification
nse_announce - 2025-11-28 18:55:36 - Appointment
nse_announce - 2025-11-28 18:13:23 - News Verification
nse_announce - 2025-11-27 14:24:12 - ESOP/ESOS/ESPS
Sources: Counter({'nse_press': 5418, 'nse_announce': 2219, 'rss': 100, 'ddg': 37})
Empty content: 0
Short content (<20 chars): 1
```



```
Merged dataset saved to: data/all_news.json
Total records: 7774
```

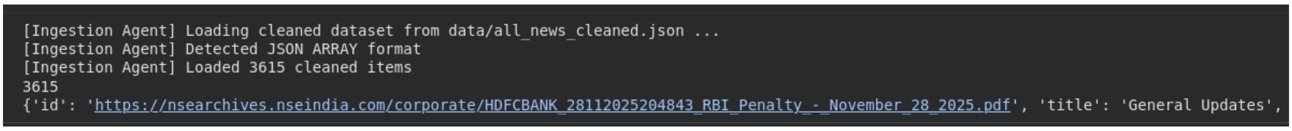
Defining Agents



Defining Multiple Agents :

- Ingestion Agent
- Duplication Agent
- Entity Agent
- Impact Agent
- Storage Agent
- Query Agent

Ingestion Agent

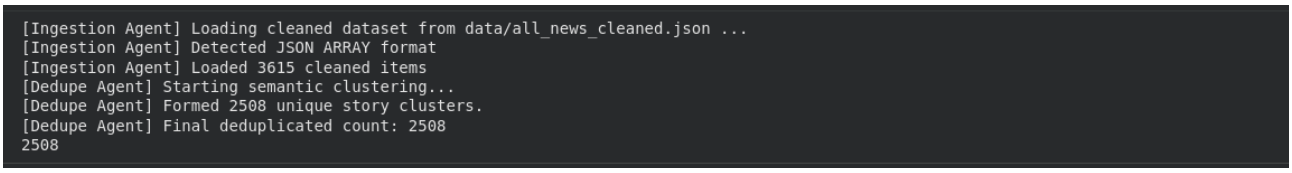


```
[Ingestion Agent] Loading cleaned dataset from data/all_news_cleaned.json ...
[Ingestion Agent] Detected JSON ARRAY format
[Ingestion Agent] Loaded 3615 cleaned items
3615
{'id': 'https://nsearchives.nseindia.com/corporate/HDFCBANK_28112025204843_RBI_Penalty_-_November_28_2025.pdf', 'title': 'General Updates',
```

Reads and loads all raw news items into memory, cleans the text, extracts metadata (title, date, source), and prepares the initial dataset for further processing. It acts as the entry point of the entire pipeline.

Duplication Agent

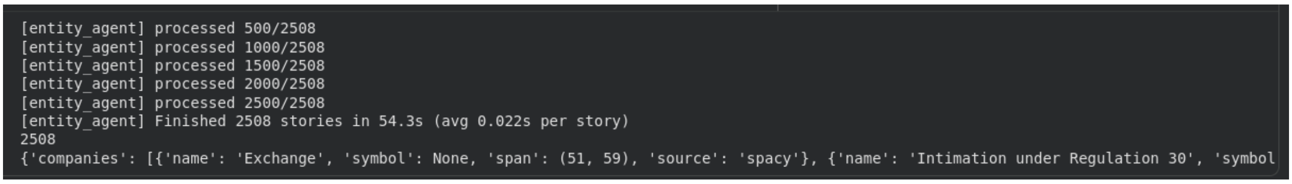
Identifies and removes duplicate or highly similar news stories using embeddings and clustering. It ensures only unique and meaningful articles are passed forward, improving accuracy and reducing noise.



```
[Ingestion Agent] Loading cleaned dataset from data/all_news_cleaned.json ...
[Ingestion Agent] Detected JSON ARRAY format
[Ingestion Agent] Loaded 3615 cleaned items
[Dedupe Agent] Starting semantic clustering...
[Dedupe Agent] Formed 2508 unique story clusters.
[Dedupe Agent] Final deduplicated count: 2508
2508
```

Entity Agent

Extracts key financial entities such as companies, regulators, money values, locations, and people using spaCy (with optional LLM assistance). This agent enriches each story with structured entity information.



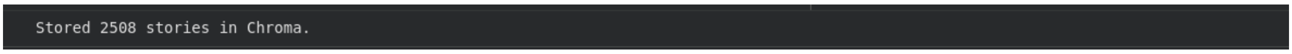
```
[entity_agent] processed 500/2508
[entity_agent] processed 1000/2508
[entity_agent] processed 1500/2508
[entity_agent] processed 2000/2508
[entity_agent] processed 2500/2508
[entity_agent] Finished 2508 stories in 54.3s (avg 0.022s per story)
2508
{'companies': [{'name': 'Exchange', 'symbol': None, 'span': (51, 59), 'source': 'spacy'}, {'name': 'Intimation under Regulation 30', 'symbol
```

Impact Agent

Analyzes the processed article and generates lightweight impact signals like which stocks or sectors might be affected. It assigns simple rule-based scores that can be used in downstream analytics.

Storage Agent

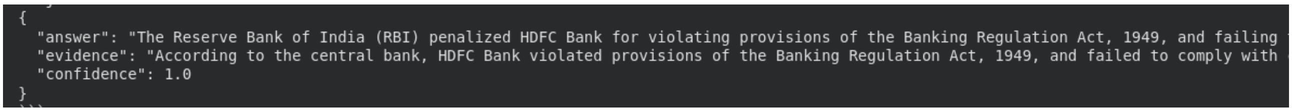
Stores the final processed stories into ChromaDB as embeddings, along with metadata. This makes the dataset searchable and enables fast retrieval during query time.



```
  Stored 2508 stories in Chroma.
```

Query Agent

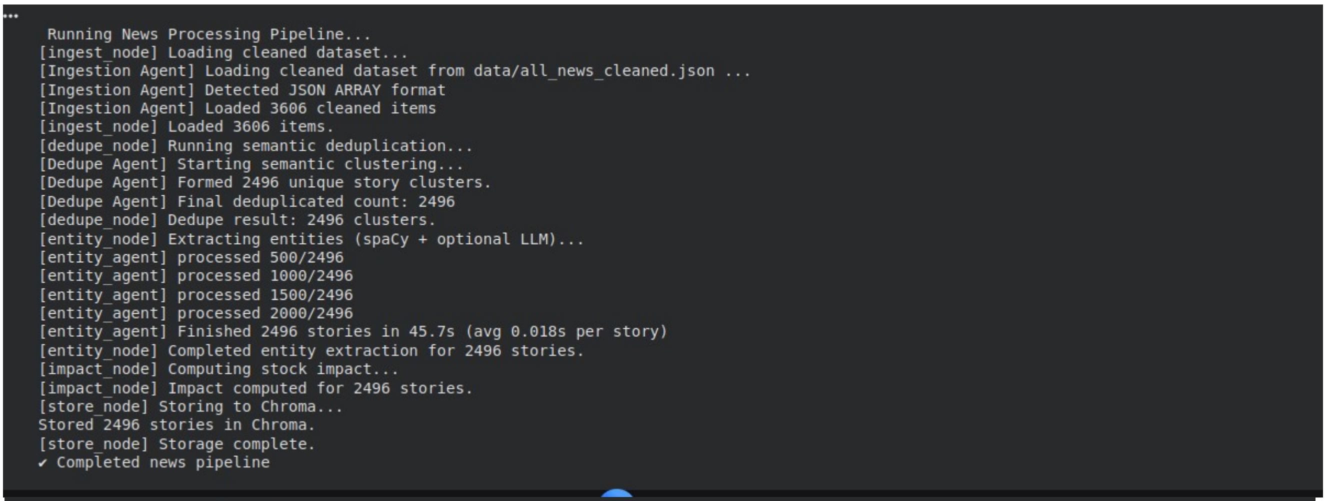
Receives a user's question, retrieves the most relevant news from ChromaDB, and generates a factual JSON-structured answer using the LLM. It ensures no hallucination and returns evidence-backed results.



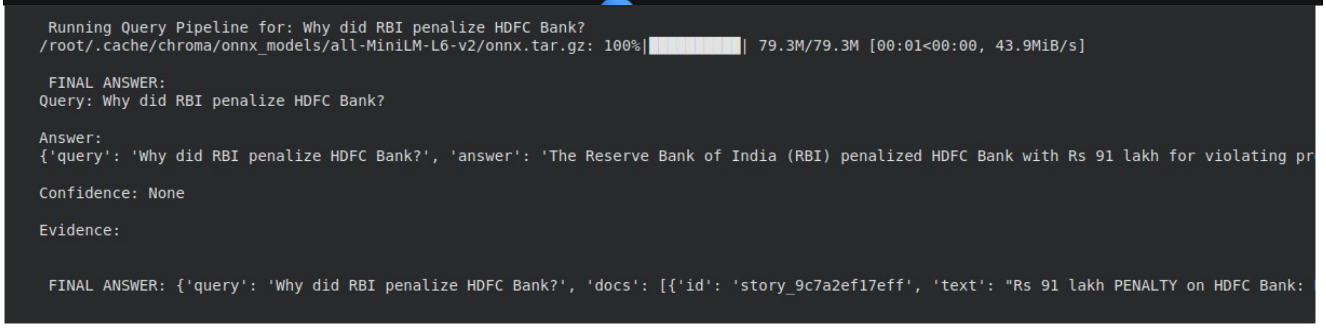
```
{
  "answer": "The Reserve Bank of India (RBI) penalized HDFC Bank for violating provisions of the Banking Regulation Act, 1949, and failing
  "evidence": "According to the central bank, HDFC Bank violated provisions of the Banking Regulation Act, 1949, and failed to comply with
  "confidence": 1.0
}
```

Multi Agent Pipeline

All agents are connected using a LangGraph-based workflow, forming a complete automated pipeline from ingestion to final query answering. Each agent performs its task independently, and the pipeline coordinates their execution in a structured flow.



```
 Running News Processing Pipeline...
[ingest_node] Loading cleaned dataset...
[Ingestion Agent] Loading cleaned dataset from data/all_news_cleaned.json ...
[Ingestion Agent] Detected JSON ARRAY format
[Ingestion Agent] Loaded 3606 cleaned items
[ingest_node] Loaded 3606 items.
[dedupe_node] Running semantic deduplication...
[Dedupe Agent] Starting semantic clustering...
[Dedupe Agent] Formed 2496 unique story clusters.
[Dedupe Agent] Final deduplicated count: 2496
[dedupe_node] Dedupe result: 2496 clusters.
[entity_node] Extracting entities (spaCy + optional LLM)...
[entity_agent] processed 500/2496
[entity_agent] processed 1000/2496
[entity_agent] processed 1500/2496
[entity_agent] processed 2000/2496
[entity_agent] Finished 2496 stories in 45.7s (avg 0.018s per story)
[entity_node] Completed entity extraction for 2496 stories.
[impact_node] Computing stock impact...
[impact_node] Impact computed for 2496 stories.
[store_node] Storing to Chroma...
Stored 2496 stories in Chroma.
[store_node] Storage complete.
✔ Completed news pipeline
```



```
 Running Query Pipeline for: Why did RBI penalize HDFC Bank?
/root/.cache/chroma/onnx_models/all-MiniLM-L6-v2/onnx.tar.gz: 100%|██████████| 79.3M/79.3M [00:01<00:00, 43.9MiB/s]

 FINAL ANSWER:
Query: Why did RBI penalize HDFC Bank?

Answer:
{'query': 'Why did RBI penalize HDFC Bank?', 'answer': 'The Reserve Bank of India (RBI) penalized HDFC Bank with Rs 91 lakh for violating pr

Confidence: None

Evidence:


 FINAL ANSWER: {'query': 'Why did RBI penalize HDFC Bank?', 'docs': [{'id': 'story_9c7a2ef17eff', 'text': "Rs 91 lakh PENALTY on HDFC Bank:
```

Response:

query: Why did RBI penalize HDFC Bank?

answer: The Reserve Bank of India (RBI) penalized HDFC Bank with Rs 91 lakh for violating provisions of the Banking Regulation Act, 1949, and failing to comply with certain RBI directions. These violations included lapses in Know Your Customer (KYC), interest rate, and outsourcing compliance.

Dataset:

press_releases.csv

Contains official press releases from financial institutions. Used for extracting raw news text, titles, and dates.

announcements_hdfc.csv

Contains HDFC related announcements and RBI-related financial news items. Used for building the core financial news corpus.

rrs_news.json

A raw dataset containing financial news articles, headlines, and announcements.

ddg_news.json

Dataset generated by your DuckDuckGo API fetcher script contains live or recent financial news scraped during testing.