PulseAI Pipeline - Final Deliverable

Team 3 | 12/09/2025

Team Members: Shin Bhide, Divyam Rana, Sidhant Sidhant, Brendan Wilcox, Huawan Zhong

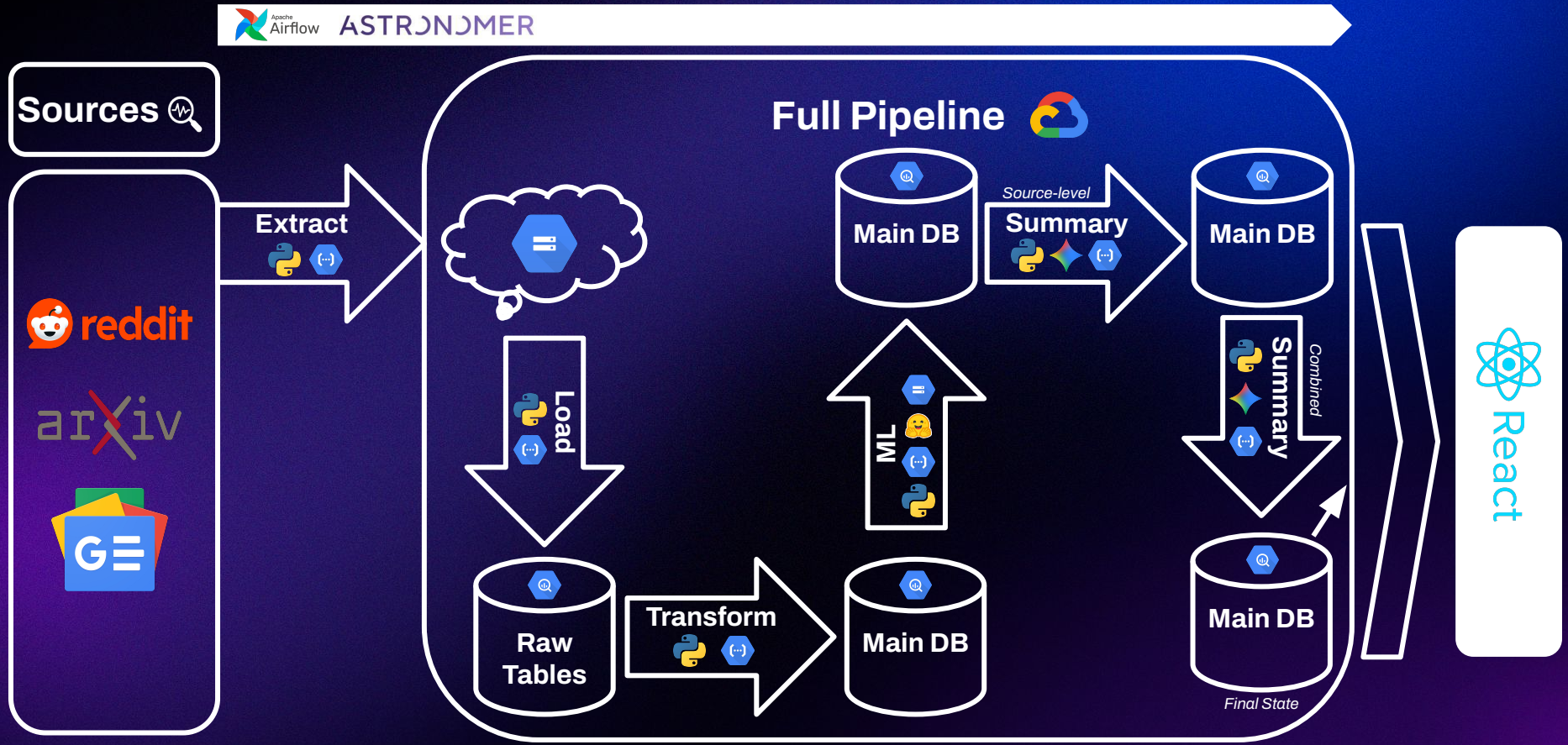
PulseAI Newsletter Demo

React

Demo

Check it out

Final Full Pipeline (High-level)



Full Orchestration

EtLT

EtL

Functionality: Ingests data from sources, light transformation, lands in “raw” tables

Schedule: Daily at 7PM

DAGs: arxiv_etl, gnews_etl, reddit_etl

Tables: papers, articles, posts

T

Functionality: More robust transformation and moves data to “main_db”

Schedule: Daily at 8PM

DAG: main_transformation

Tables: arxiv_papers, news_articles, reddit_posts, *dimension tables for each source*

ML

Functionality: Groups individual content into 1 of 9 tags

Schedule: Daily at 9PM

DAGs:
arxiv_paper_tagging, news_article_tagging, reddit_post_tagging

Tables:
arxiv_papers_tagged, news_articles_tagged, reddit_posts_tagged

LLM

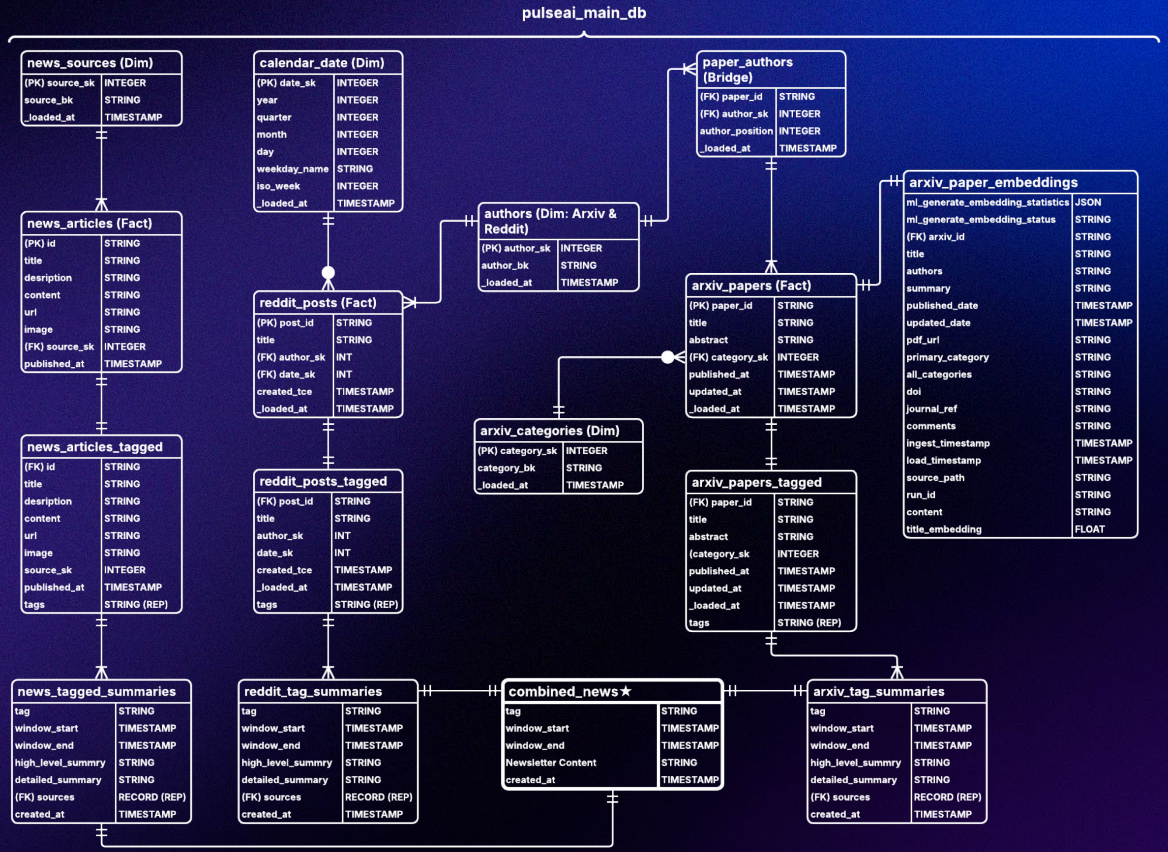
Functionality: Source-level summaries by tag → Aggregate source summaries and create single summary by tag

Schedule: Saturdays at 12AM and 3AM

DAGs: newsletters_weekly, combined_newsletter

Tables:
arxiv_tag_summaries, news_tagged_summaries, reddit_tag_summaries, combined_newsletter

Final Data Model

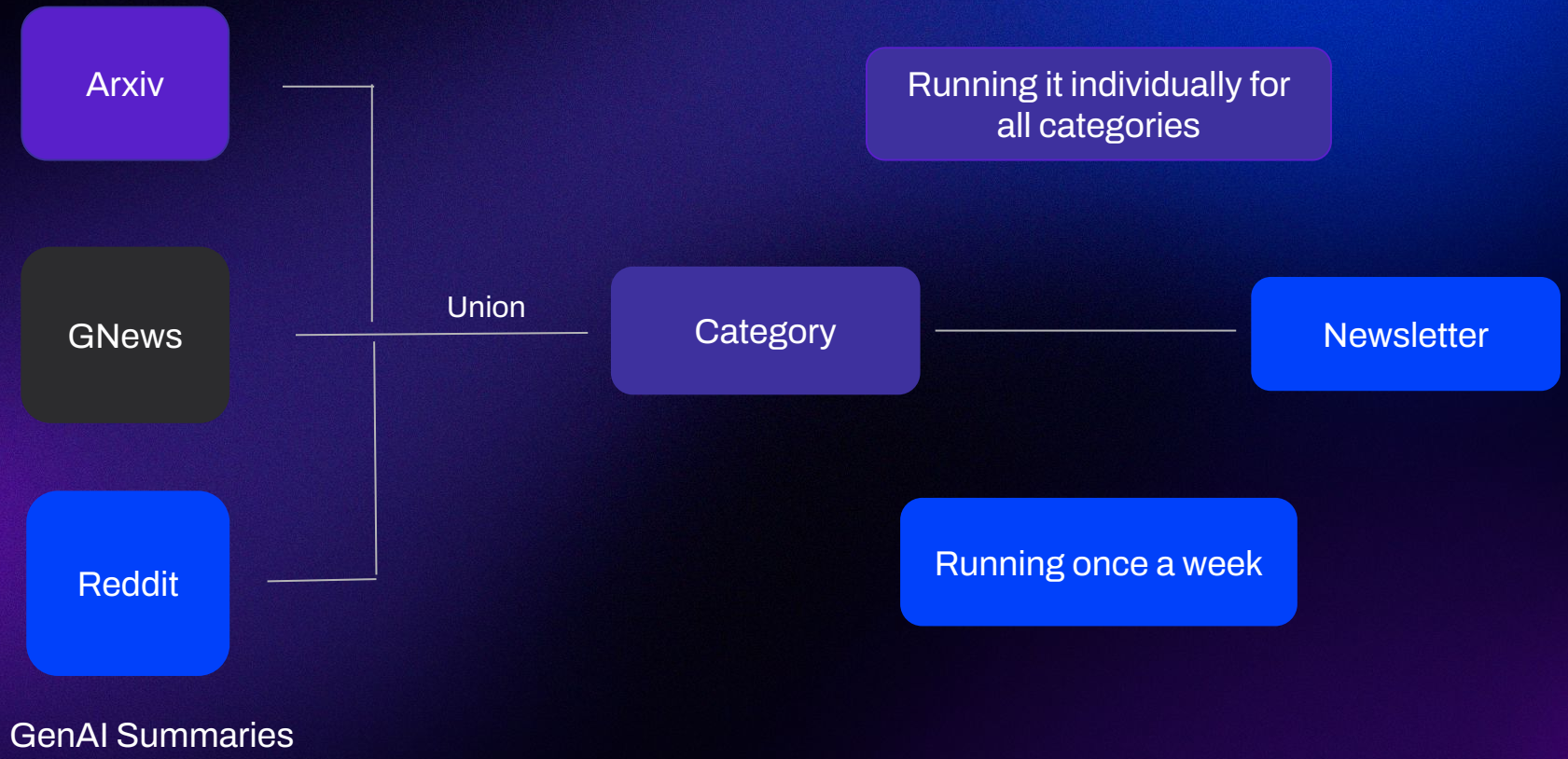


Hugging Face Update

- Used a zero-shot Hugging Face MNLI model to tag articles by topic.
- The Classifier run by the Cloud Function is cached in the GCS bucket
- No fine-tuning needed since the model already works well on mixed text
- Flexible setup that can support more categories in the future

Generative AI Workflows

Generative AI Workflows



Semantic Search Functionality

- Vector database embedding model in BigQuery
- Converts title into embeddings, cosine similarity with user query
- Handled serverlessly with BigQuery SQL and Vertex AI
- e.g. “Gogle” will map to Google

Gen AI Prompt Evaluation

- LLMs-as-judge: ‘Pro’ version judges ‘fast’ version
- List-wise execution: all prompt outputs judged together in one go
- Hallucination score (lowest is best)
- Completeness and Conciseness score (highest is best)
- Human-in-the-loop for final validation

Gen AI Prompt Evaluation

Hallucination score (lowest is best)

	Prompt index	Source 1 (Physics)	Source 2 (History)	Source 3 (Finance)	Average
0	101	1.000000	1.000000	1.500000	1.166667
1	102	1.500000	1.500000	2.000000	1.666667
2	103	4.000000	1.000000	1.000000	2.000000

Completeness and conciseness score (highest is best)

	Prompt index	Source 1 (Physics)	Source 2 (History)	Source 3 (Finance)	Average
0	101	9.000000	8.500000	9.000000	8.833333
1	102	7.000000	7.000000	8.500000	7.500000
2	103	8.000000	9.000000	7.000000	8.000000

Conclusion & Practical Implications

- Use Airflow to orchestrate all Cloud Functions and ML jobs in one place
- Ingested articles/posts daily from Arxiv/Reddit/GNews automatically
- Tracked volumes and trends over time for 9 topics
- Combined three sources and generated detailed summaries along with category tags
- React App supports search, basic downloads, and exploratory visualizations

Challenges & Next Steps

- The zero-shot classifier’s still under detects the healthcare topic
- Hallucinations still occur in the Gen AI summaries
- Ingest more data from current sources
- Add more sources and give the model more context around it

Thank you | Q&A