PulseAI Pipeline - Phase 2

Team 3 | 11/11/2025

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

Agenda

01

Project Overview

02

Updated Data Model

03

Updated Pipeline

04

Machine Learning

05

Streamlit

06

Looking Ahead

07

Q&A

Project Overview: AI-Curated Newsletter

Motivation

- **AI research** and industry **news** are **scattered**, **fast-evolving**, and **difficult to track**
- A **unified data pipeline** captures diverse sources to **deliver timely, reliable insights** via an **AI/ML-backed workflow**

Data Overview

- **Sources:** reddit arXiv Google News
- Data will be **ingested daily**, aligned with the newsletter publishing cycle

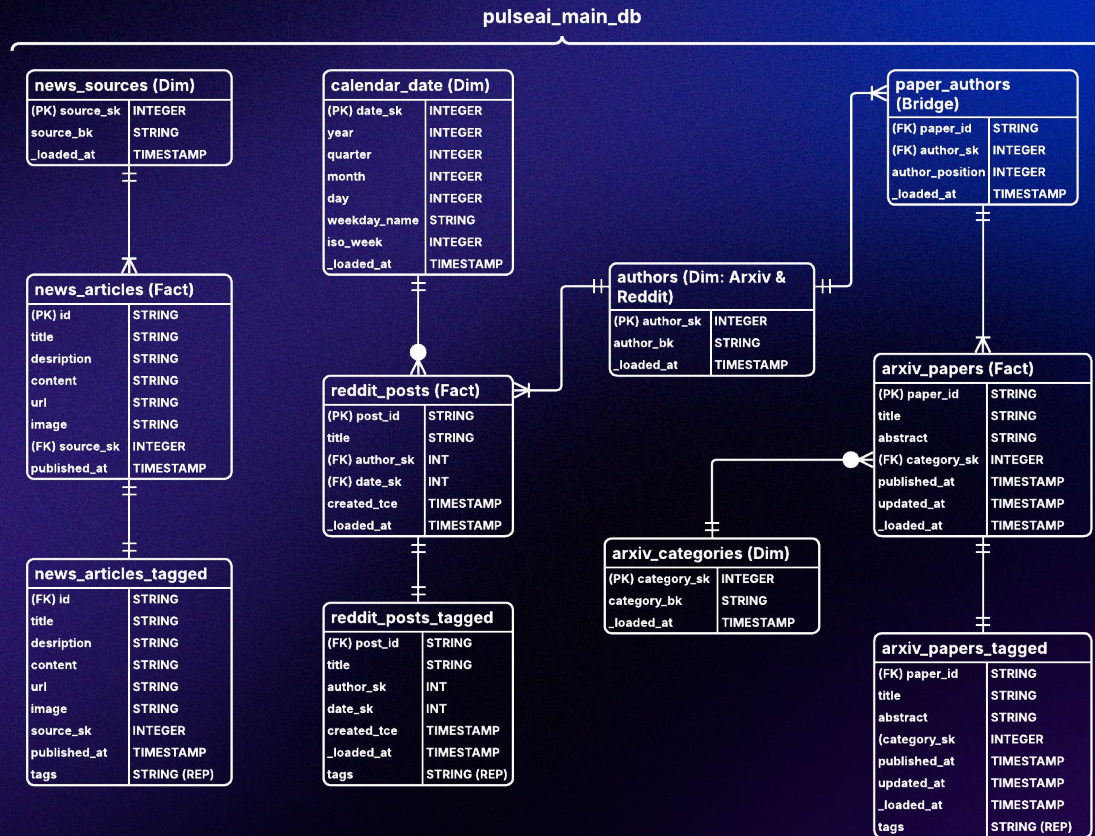## Problems Explored

- Information overload, trend identification, decision support, and using AI to automate data collection, classification, and summarization
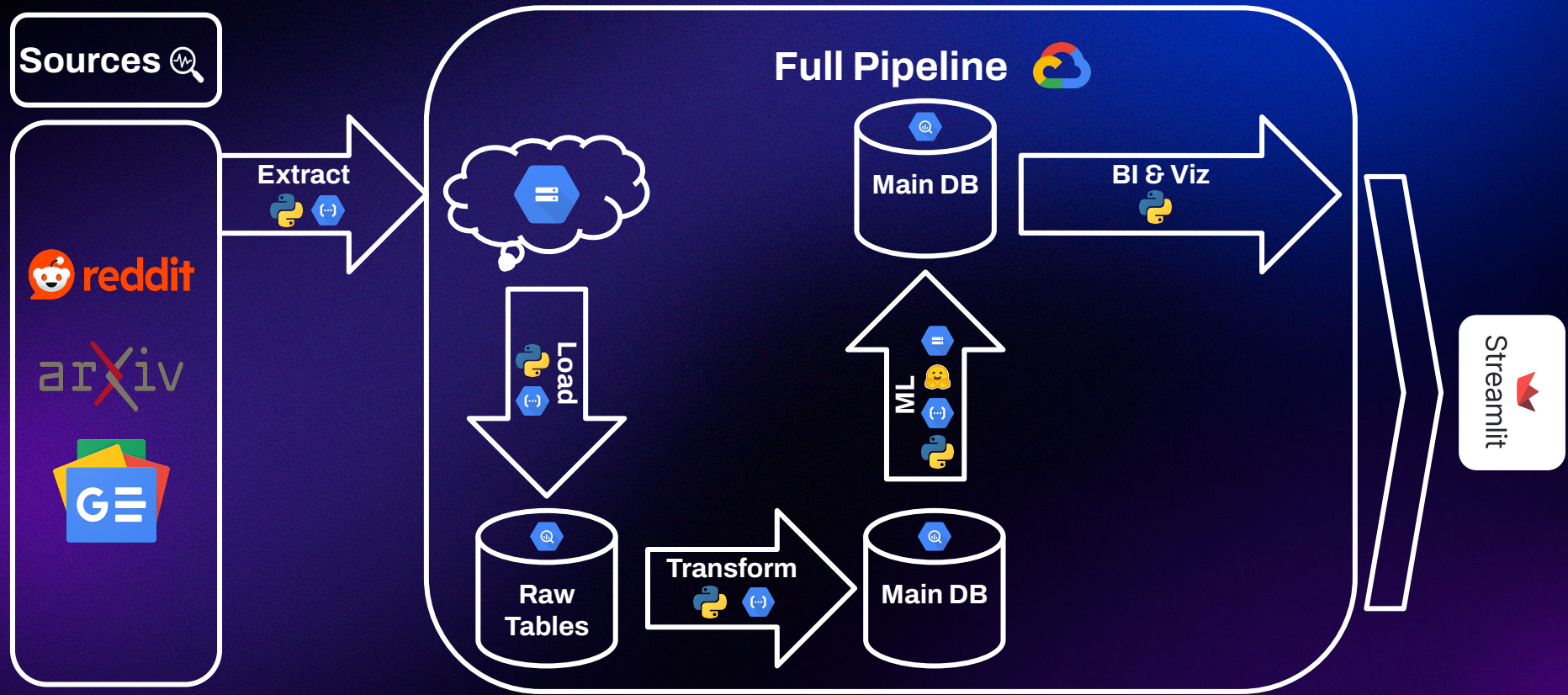
Final Output

- Weekly AI-curated **newsletter** backed by an analytics **dashboard**
- Delivers **concise summaries** with an **analytics dashboard** tracking key metrics, offering a **reliable**, **time-saving view** of AI developments

Updated Data Model

pulseai_main_db

news_sources (Dim)

(PK) source_sk	INTEGER
source_bk	STRING
_loaded_at	TIMESTAMP

calendar_date (Dim)

(PK) date_sk	INTEGER
year	INTEGER
quarter	INTEGER
month	INTEGER
day	INTEGER
weekday_name	STRING
iso_week	INTEGER
_loaded_at	TIMESTAMP

paper_authors (Bridge)

(FK) paper_id	STRING
(FK) author_sk	INTEGER
author_position	INTEGER
_loaded_at	TIMESTAMP

news_articles (Fact)

(PK) id	STRING
title	STRING
desription	STRING
content	STRING
url	STRING
image	STRING
(FK) source_sk	INTEGER
published_at	TIMESTAMP

authors (Dim: Arxiv & Reddit)

(PK) author_sk	INTEGER
author_bk	STRING
_loaded_at	TIMESTAMP

arxiv_papers (Fact)

(PK) paper_id	STRING
title	STRING
abstract	STRING
(FK) category_sk	INTEGER
published_at	TIMESTAMP
updated_at	TIMESTAMP
_loaded_at	TIMESTAMP

reddit_posts (Fact)

(PK) post_id	STRING
title	STRING
(FK) author_sk	INT
(FK) date_sk	INT
created_tce	TIMESTAMP
_loaded_at	TIMESTAMP

arxiv_categories (Dim)

(PK) category_sk	INTEGER
category_bk	STRING
_loaded_at	TIMESTAMP

news_articles_tagged

(FK) id	STRING
title	STRING
desription	STRING
content	STRING
url	STRING
image	STRING
source_sk	INTEGER
published_at	TIMESTAMP
tags	STRING (REP)

reddit_posts_tagged

(FK) post_id	STRING
title	STRING
author_sk	INT
date_sk	INT
created_tce	TIMESTAMP
_loaded_at	TIMESTAMP
tags	STRING (REP)

arxiv_papers_tagged

(FK) paper_id	STRING
title	STRING
abstract	STRING
(category_sk	INTEGER
published_at	TIMESTAMP
updated_at	TIMESTAMP
_loaded_at	TIMESTAMP
tags	STRING (REP)

Updated Full Pipeline

Machine Learning – Basic Topic Modeling

Gnews (Non-technical Audience)

- Legal rulings & IP disputes
- Big Tech & AI industry
- AR/VR hardware & optics
- Music, streaming & copyright
- Middle East, oil & funds
- Transport safety & fleets
- US–China–Korea geopolitics, chips
- Tech misuse & education issues
- Travel tech & hospitality
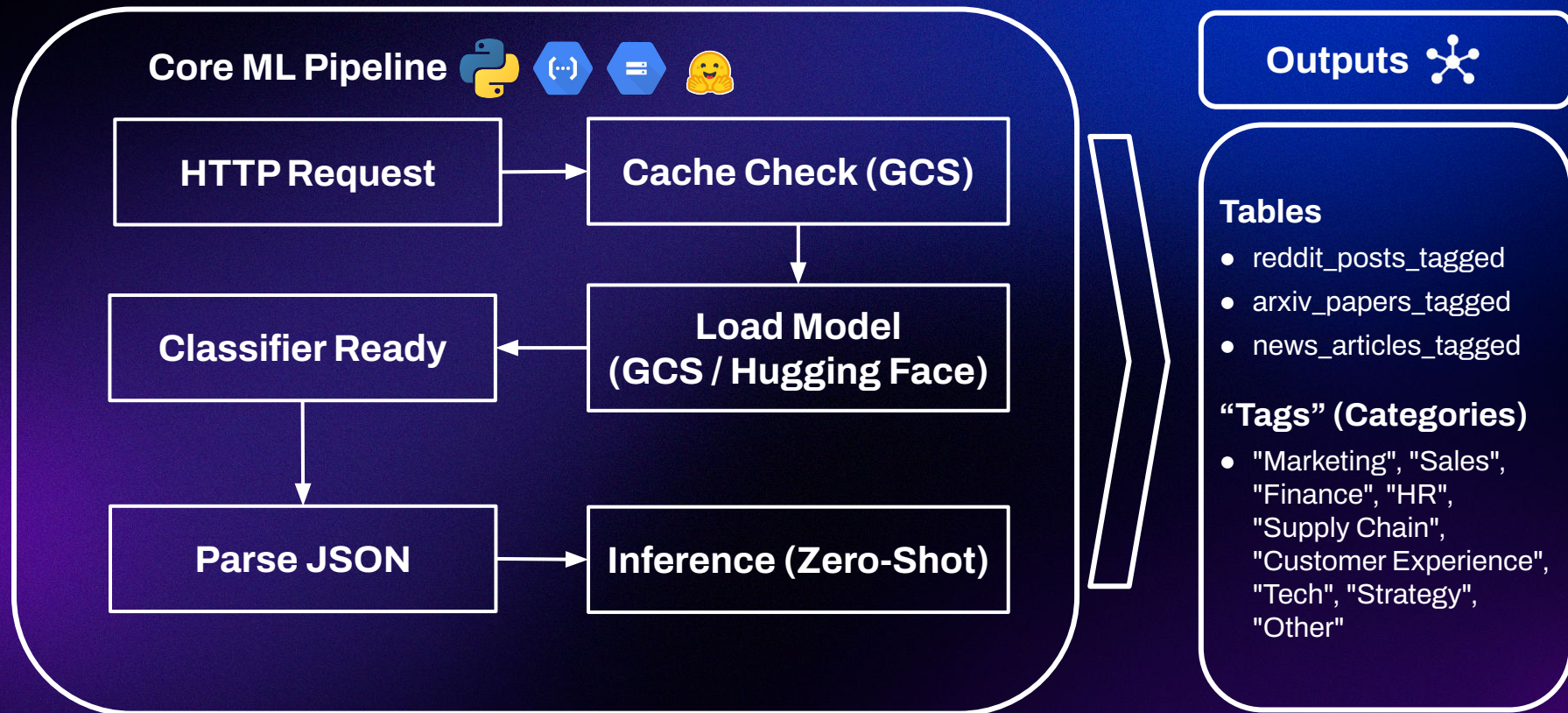
Arxiv & Reddit (Technical Audience)

- Machine Learning Model & Training
- Theoretical & Mathematical ML Foundations
- French-language text
- Software Tools & API/Data Process
- AI Agents & Game Simulations
- AI in Medical Imaging
- Deep Learning/ Neural Representation Theory

Machine Learning – Hugging Face

- **Goal:** Standardize topic labels across sources

- **Method:** Zero-shot classifier deployed via Cloud Function

- **Model:** distilbart-mnli via transformers uploaded on GCP bucket

- **Output:** 9 business topics – Marketing, Sales, Finance, HR, Supply Chain, Customer Experience, Tech, Strategy, and Other.

- **Why it matters:** Powers cross-source analytics & dashboard

ML Pipeline Visualized

Core ML Pipeline 🐍 🔵 🟦 🤗

HTTP Request → **Cache Check (GCS)**

Cache Check (GCS) → **Load Model (GCS / Hugging Face)**

Load Model (GCS / Hugging Face) → **Classifier Ready**

Classifier Ready → **Parse JSON**

Parse JSON → **Inference (Zero-Shot)**

Outputs ✴️

Tables
- reddit_posts_tagged
- arxiv_papers_tagged
- news_articles_tagged

"Tags" (Categories)
- "Marketing", "Sales", "Finance", "HR", "Supply Chain", "Customer Experience", "Tech", "Strategy", "Other"

Looking Ahead

- Improving UI and UX of the application

- Gen AI integration to summarize latest developments

- Improving ML models for better topic modelling

- Further database standardization and optimization - Naming conventions, dropping unused columns, etc.

Thank you | Q&A