PulseAI Pipeline - Phase 1

Team 3 | 10/21/2025

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

Agenda

01

Project Overview

02

Data Model

03

Pipeline Flow

04

Orchestration Design

05

BigQuery Usage

06

Initial Visualization

07

Looking Ahead

08

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

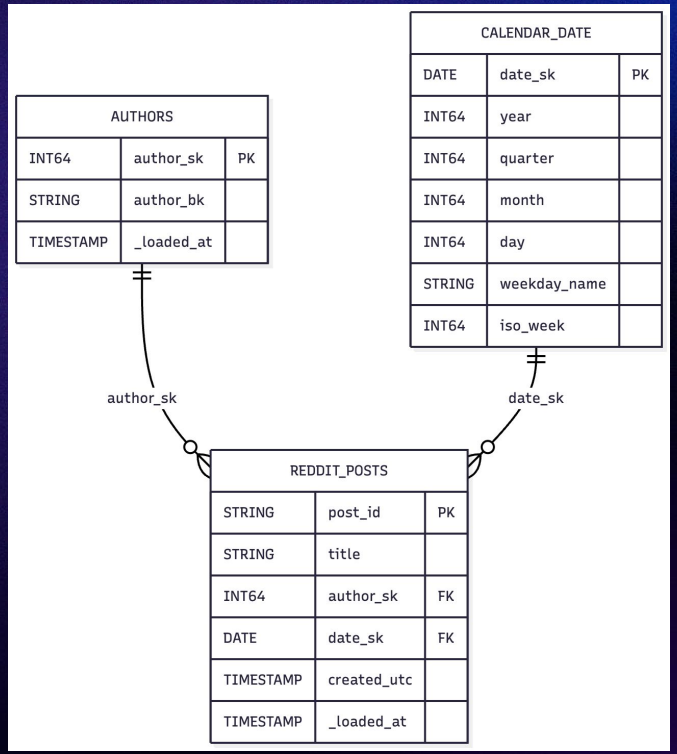
- Information overload, trend identification, decision support, and using AI to automate data collection, classification, and summarization

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

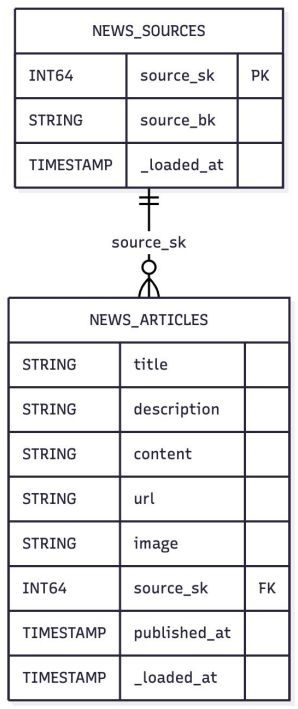
Data Model

REDDIT

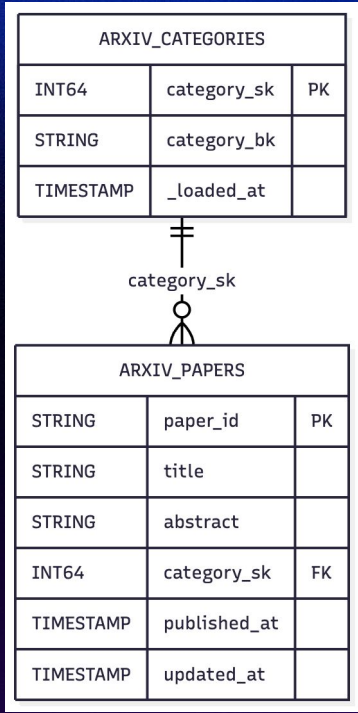


Data Model

GNews



ARXIV



Pipeline Flow

EtLT process to establish a secure, scheduled, and scalable ingestion process for raw data.

1. Execution Environment (GCF): Core API fetching and data acquisition logic was wrapped into scalable Google Cloud Functions (GCF), serving as the lightweight, serverless execution platform.

2. Credential Security: Sensitive API keys were centrally managed to ensure secure storage and runtime retrieval of credentials by GCFs, thereby minimizing security risks.

3. Staging Layer: Raw, immutable output files are persistently stored in GCP Buckets as a Data Lake, providing a reliable staging area before transformation.

Orchestration, Scheduling, & Data Readiness

1. Pipeline Orchestration: Astronomer (Apache Airflow) DAGs were utilized to manage dependencies, control the flow, and automate the pipeline execution.

2. Cron Scheduling: The DAGs enforce a fixed cron schedule (twice daily) for the automatic, periodic execution of the Cloud Function extraction tasks.

3. Data Integration (BigQuery): Ingested raw data was integrated into BigQuery, optimizing it for high-performance SQL querying and serving as the primary source of truth.

4. ML/NLP Readiness: The structured BigQuery data is ready for downstream use in Natural Language Processing (NLP) and Machine Learning (ML) tasks and dashboards.

5. Quality Assurance: Rigorous testing was performed to validate the pipeline's functionality and resilience under various conditions, ensuring reliable operation.

BigQuery & GCP: Capabilities & Intended Use

Ingestion

- Utilize **“extract” *Cloud Functions*** that source data using **API calls**
- Lands **raw data in GCS**

Loading

- **“Setup” *Cloud Function*** creates **dataset, table, and schema** expected by load function
- **“Load” *Cloud Function*** Deduplicates extracted data and **loads into BQ**

Transform

- Use **SQL queries** to clean, join, and enrich data as needed
- Create **views** that **feed dashboard** and **AI/ML** workflows

Analytics, Reporting

- Diverse **compatibility with most BI tools**
- Can utilize **Dataproc** and **Vertex AI** to handle AI/ML
- Can **optimize** for **software capabilities** and **team preference**

Orchestration & Design Choices

ARXIV Pipeline

Extracting the research paper data from the arxiv database regarding ML and AI development

Gnews Pipeline

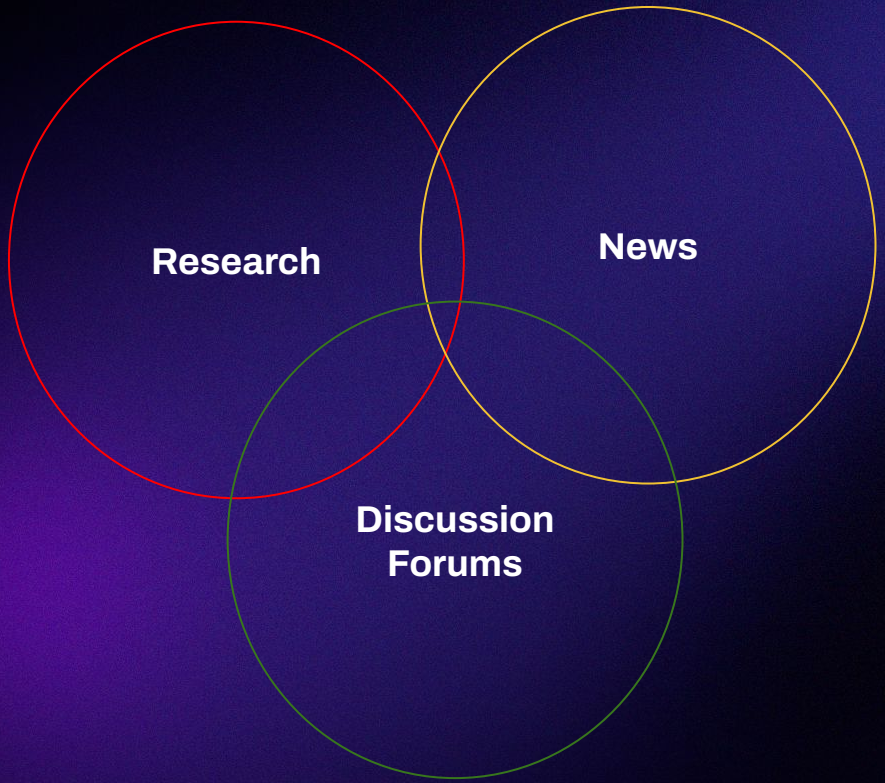
Fetching the latest news in the AI and ML field to keep up with developments

Reddit Pipeline

Looking for discussions in popular ML and related subreddits to keep up with the discussions

AI Newsletter

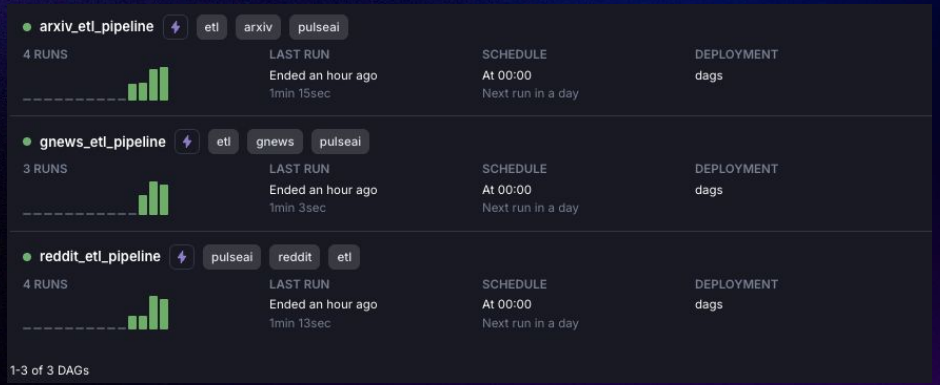
Orchestration & Design Choices



Setup Bigquery table

Extract data thru APIs + Scraping

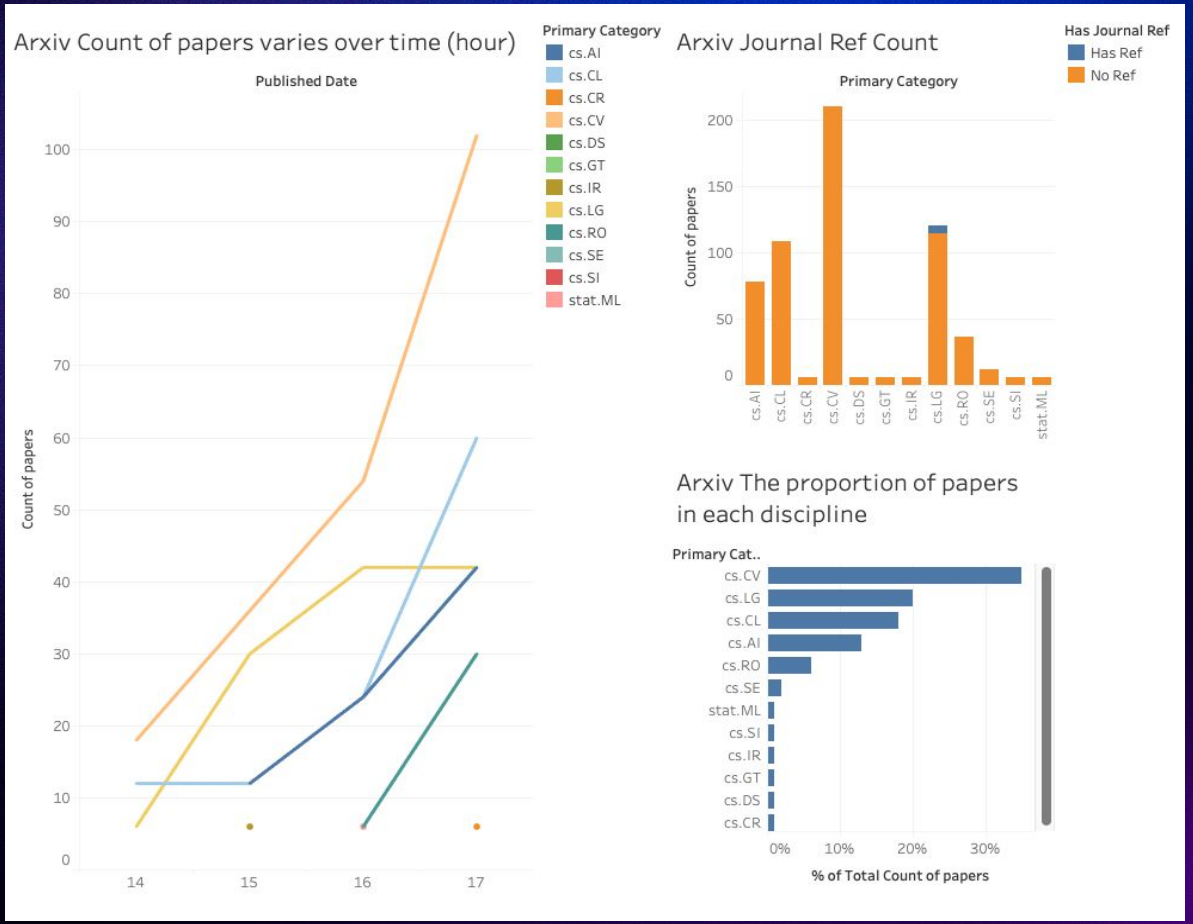
Load to bigquery from GCP Bucket



Initial Visualizations

Some Basic Insight for Arxiv:

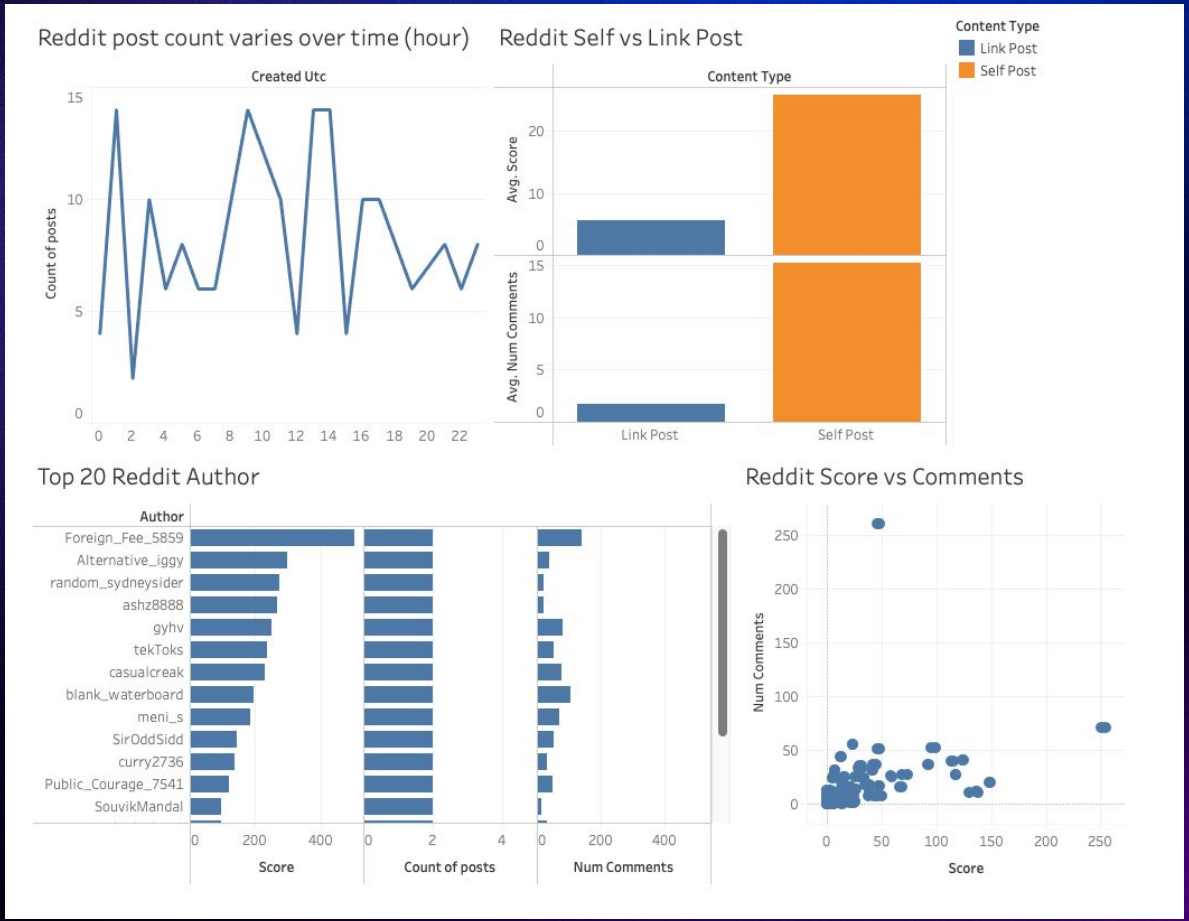
1. Most paper are published around 5pm
2. Only a small fraction of papers include a Journal reference
3. Computer Science dominates the dataset
4. The top three disciplines account for the majority of papers



Initial Visualizations

Some Basic Insight for Reddit:

1. Post volume varies a lot by hour
2. Self posts get much higher scores and more comments than link posts
3. Most of top authors have 2 posts in reddit
4. Positive but weak relationships between score and comments



Looking Ahead

- **Automated Content Classification:** Use NLP/ML to auto-categorize (into topics e.g. Healthcare, New Gen AI) all incoming data.
- **Unified Data Model:** Standardize content into a single, classified data model.
- **GenAI Newsletter Generation:** Implement Generative AI to automatically create newsletters from categorized content

Thank you | Q&A