

AI 620 Fundamentals of Data Engineering

Assignment 1 Report

Name: Noor UI Ain Anwar

Student ID: 25280062

GITHUB REPOSITORY URL

1. Thematic Focus: AI Labor Markets

In this assignment, I chose to work in the thematic area of AI Labor Markets. This topic explores how the employment landscape in Artificial Intelligence changes, and in particular:

- Job Trends: What jobs (e.g., Data Scientist vs. AI Engineer) are in high demand?
- Skill Demands: What specific technical skills are appearing most frequently in news and job descriptions?
- Economic Effect: The relationship between the searches on AI terms and salary standards.

In order to examine this, I created a modular Extract-Load-Transform (ELT) pipeline that combined three different data sources:

1. NewsAPI (Unstructured): Real-time articles on "AI jobs" and "tech layoffs."
2. Kaggle (Structured): The *Global AI & Data Science Job Market (2020-2026)* dataset containing 50,000+ salary records.
3. Google Trends (Semi-Structured): 5-year historical search interest data for terms like "Generative AI" and "Machine Learning."

2. Key Findings & Insights

When the transformation and visualization layers of the pipeline were run, three different patterns were found:

- Temporal Explosion of Generative AI: Temporal analysis of search rates showed that the interest in the concept of Generative AI increased exponentially and suddenly in the end of 2022. This is closely associated with the public launch of ChatGPT implying that the market shifted to a new trend whereby "Generative AI" started taking over "Data Science" in popular attention.
- Role Dominance: Categorical analysis of the Kaggle data revealed that Data Engineer and Data Scientist are the most updated postings. The underlying

infrastructure jobs (engineering) continue to be the main drivers of hire despite the hype about the existence of AI Research.

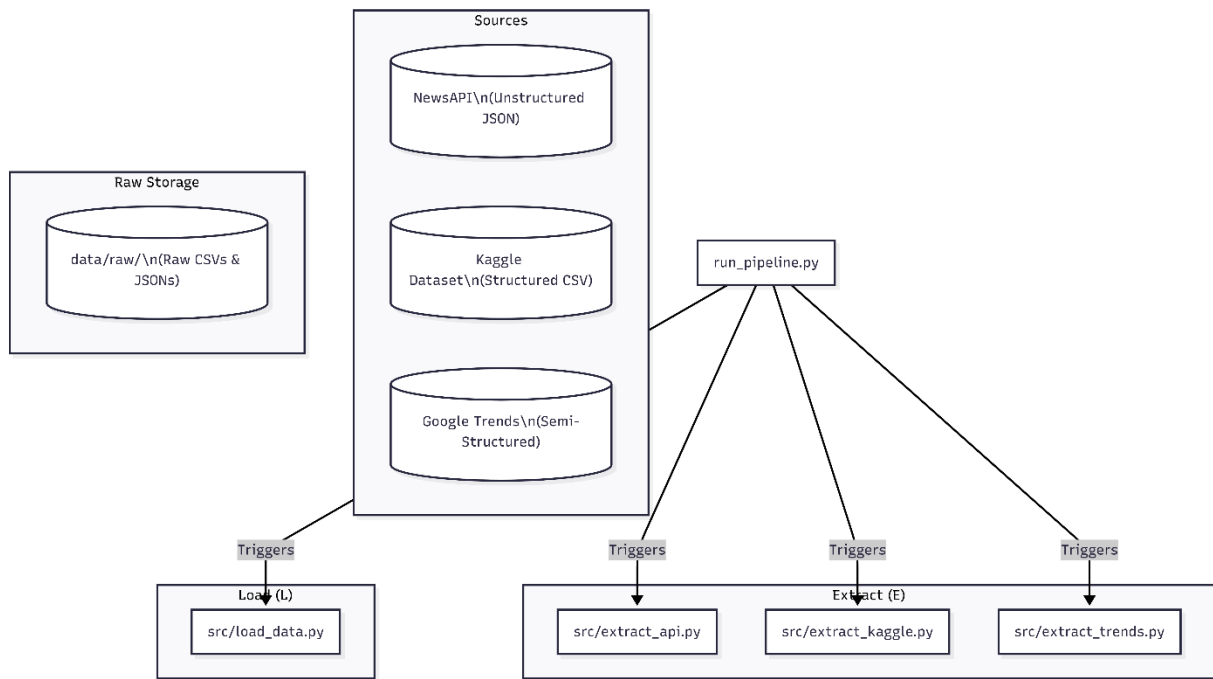
- **Experience-based Salary Variance:** The boxplot analysis reported high positive correlation between experience level and salary, with the Executive (EX) roles exhibiting much greater variance than the Entry-level (EN) positions, which means that the pay to the top-ranking senior staff member is much more specific to the niche services or the size of the company, whereas the compensation to the staff ID is more standardized.

3. Technical Challenges Encountered

Implementing the pipeline revealed several data engineering hurdles:

- **Authentication & Security:** Managing API credentials securely was a primary challenge. The Kaggle API initially failed to authenticate via the standard `kaggle.json` method due to browser download issues. I resolved this by implementing a robust `dotenv` configuration to load credentials directly from environment variables, ensuring the pipeline remains secure and portable.
- **Data Heterogeneity:** Ingesting data from disparate sources required careful handling. The Google Trends data arrived with metadata columns (e.g., `isPartial`) that needed pruning, while the Kaggle dataset contained duplicate rows. Standardizing these into a unified processed/ format (CSV for analysis, JSON for lineage) was critical for downstream compatibility.
- **Rate Limiting:** The Google Trends API (`pytrends`) is sensitive to frequent requests, occasionally returning 429 Too Many Requests errors. I addressed this by implementing `try-except` blocks to handle connection failures gracefully and advising a cool-down period between pipeline runs.

4. Pipeline Diagram



5. Conclusion

This ELT pipeline successfully demonstrates how heterogeneous data sources can be unified to derive actionable insights. By automating the extraction and cleaning of labor market data, this system provides a foundation for monitoring the real-time evolution of the AI workforce.