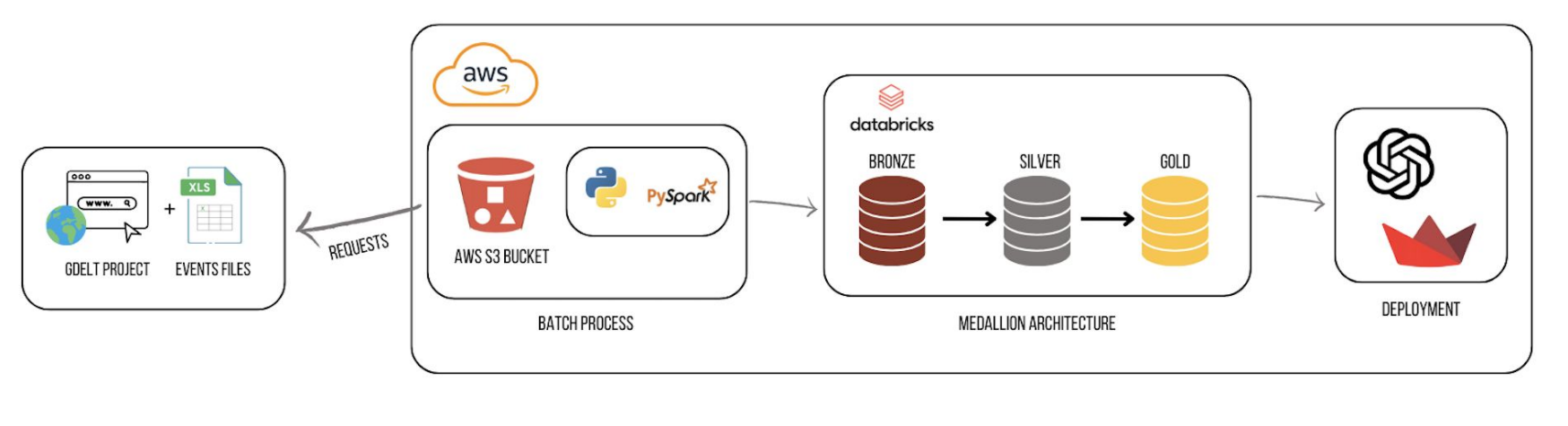
Factored Datathon 2024
Exploding Gradients

Technical Details

- **Data Storage and Management**: Utilized AWS S3 for scalable storage of raw GDELT event data, and employed Delta tables in Databricks for managing data through the Medallion architecture, ensuring consistency and reliability.
- **Data Processing Pipeline**: Implemented a three-layer data processing pipeline:
  - Bronze: Ingested raw data from AWS S3 into Databricks, preserving all unprocessed data.
  - Silver: Cleaned and aggregated data from the Bronze table, removing duplicates and computing metrics like Average Number of Sources and Average Tone.
  - Gold: Enriched data by calculating an importance metric for each news event
- **Automated Data Pipeline**: Implemented a Databricks job to scrape daily data from GDELT, upload to AWS S3, and process through the Bronze, Silver, and Gold layers, ensuring up-to-date information for analysis.
- **Data Analysis and Machine Learning**: Employed various methods including normalization, weighted scoring for news ranking, LDA for topic classification, and GPT API for high-quality, abstractive article summarization.
- **Web Application Deployment**: Developed a Streamlit web app hosted on Streamlit Cloud, integrated with Databricks for real-time data access, allowing users to select date ranges and categories and view top news summaries.

Data Pipeline



Final Tool

News Categories

Select categories to display

All

Top News Articles Around the World

Current date: 26/08/2024 03:05

Get news from:

- Yesterday
- Choose dates

Start date: 2024/08/14

End date: 2024/08/17

Number of news articles to display: 5

Fetching news from 2024-08-14 to 2024-08-17

- Uncovering torture, sexual assault in 'Israel's Guantanamo Bay' prison
- Authorities say pharmacist will plead no contest to involuntary manslaughter in 11 Michigan deaths tied to bad steroids
- Jury convicts White Florida woman in fatal shooting of Black neighbor during ongoing dispute
- Lawyers push for massive $58 billion compensation for Ethiopian Airlines crash victims' families

Impact and Importance

- **Enhanced Decision-Making:** Delivers precise news summaries to empower timely, well-informed decisions.
- **Improved Data Quality**: Ensures high-quality, relevant information through rigorous data processing and filtering.
- **User-Friendly Access:** Offers an intuitive platform for customized, interactive news exploration.
- **Impact-Driven Insights**: Utilizes importance-based metrics to rank and highlight the most significant news events, making it easier to identify and act on high-impact information.
- **Commercial and Strategic Value**: It serves as a valuable tool for monitoring trends and responding to changes in the market or geopolitical landscape.

Tool Usage Examples

- **Media Companies:** Use the tool to identify trending topics and guide content creation strategies. Analyze news summaries to tailor content to audience interests and improve engagement.
- **Financial Institutions:** Monitor news summaries to forecast market trends and make informed investment decisions. Evaluate news sentiment and importance to assess potential risks to investments or financial stability.
- **Government Agencies:** Utilize news summaries to understand public sentiment and inform policy development. Quickly assess the impact of current events to coordinate responses and allocate resources efficiently
- **Retail Companies:** Adjust marketing strategies and product offerings based on news about consumer trends and economic conditions.
- **Non-Governmental Organizations (NGOs):** Support advocacy campaigns by tracking and summarizing news on social issues.