Build a Real-Time RAG App with Pathway

1. Background & Industry Context

Generative AI is rapidly evolving from a niche tool into a core enterprise technology. Gartner predicts that as organizations move from pilot projects to production-ready implementations, Retrieval-Augmented Generation (RAG) will become central to AI strategies. RAG enhances language models like ChatGPT by:

- Retrieving relevant information from large, dynamic corpora.
- Feeding that context into Large Language Models to generate more accurate answers—even for data that wasn't part of the original training.

What is Agentic RAG? *(Optional)*

Agentic RAG adds autonomy to standard RAG systems. Instead of manually orchestrating each retrieval step, an "agent" (your AI application) can:

- Decide how to retrieve information,
- Manage retrieved data,
- Apply intelligent strategies (multi-agent collaboration, corrective RAG, etc.)

2. About Pathway

About the Company: Pathway Technology Inc. (pathway.com) is the maker of the world's fastest global data processing engine (GitHub). With offices in the US, France, and Poland, our ~25-member team has deep expertise from top AI labs like Microsoft Research, Google Brain, and ETH Zurich. Many of our members have worked at Google and hold degrees from prestigious institutions like École Polytechnique, UC Berkeley, CNRS, and HEC Paris—one even earned a PhD at just 20. Our CTO has co-authored notable works with AI pioneers Geoffrey Hinton and Yoshua Bengio. Our leadership includes the co-founder of Spoj.com (one of the earliest CP platforms with over 1M developers) and NK.pl (Poland's first social media with 13.5M+ users), and advisors from current/previous leadership of OpenAI, SAP, and DHL.

Join our open-source community by starring the GitHub repositories below.

- https://github.com/pathwaycom/llm-app
- https://github.com/pathwaycom/pathway

About Pathway Framework

Pathway is a Python data processing framework designed for analytics and AI pipelines over data streams. It is the ideal solution for real-time processing use cases such as streaming ETL or Retrieval Augmented Generation (RAG) pipelines for unstructured, changing data.

Key Components of this definition:

1. **Python Framework:** Written in Rust 🦀 for speed and efficiency, Pathway is usable via Python, making it powerful yet simple to use with just Python know-how.
2. **Data Processing:** Pathway excels at processing large-scale, real-time data and is recognized as the world's fastest data processing engine. As a developer, you can use it for tasks like performing JOINS on incoming data streams (real-time data flow) or updating vector/hybrid indexes in real time. These are just simple examples—its potential goes much further.
3. **AI Pipelines Over Data Streams:** Pathway helps AI systems learn from real-time data streams, enabling applications like sentiment analysis, anomaly detection, and RAG pipelines that automatically adapt to incoming data.

3. Problem Statement

Title: *Build a Real-Time Retrieval-Augmented Generation (RAG) App with Pathway*

At Pathway, we offer several app templates with end-to-end executable code in Dockerized environments. However, for this challenge, we expect participants to explore and build solutions using their own expertise, rather than relying on a pre-defined template.

Objective:
Create a fully functional end to end real-time RAG application that leverages **Pathway** as its core orchestrator for data ingestion, incremental indexing (vector/hybrid), and REST API deployment. You are free to use any domain (health, finance, etc.) and any agentic framework (e.g., LangGraph, Crew AI, AutoGen, OpenAI Swarm) *if* you want to add agent functionality. However, your solution **must** demonstrate how Pathway's real-time pipeline automatically adapts to data updates and provides fresh context to the LLM or agent.

Minimum Requirements

1. **Pathway for Real-Time Data Sync**
    - Ingest data from at least one real-time data source (e.g., a streaming API, file system changes, or database updates). You can connect to sources like Google Drive, APIs or simulated real-time ingestion via JSONLines or other data sources that you deem necessary.
    - Show how Pathway continuously indexes or re-indexes this data in near real time.
    - If there's a real-world use-case and you don't find an API, you can also write a script to add or modify a doc every few minutes, demonstrating your app's ability to automatically re-index whenever a change is made.
2. **Vector Store / Document Store Setup**
    - Use Pathway's built-in or custom vector/hybrid indexing approach to store and retrieve documents.
3. **RAG Pipeline**
    - Implement a retrieval-augmented generation pipeline that leverages Pathway's indexing for context retrieval.

    - The LLM queries must be answered with references to the newly ingested documents to show real-time updates.
4. **Deployment via Pathway's REST API Endpoint**
    - Expose the RAG pipeline (and optionally your agent logic) as a REST API endpoint deployed through Pathway.
    - The endpoint should accept queries and return answers enriched with context from the real-time data store.
5. **Basic UI**
    - Provide a minimal user interface (it can be a simple web app or CLI) to demonstrate querying the RAG endpoint.
    - The UI does not need to be fancy, but it should allow easy testing of real-time updates.
6. **Deliverables**
    - **GitHub Repo**: Must contain all code (Python scripts, YAML/SQL configuration (not required if working in python), Dockerfiles, etc.) and instructions for setup.
    - **Documentation**: Clear README with setup steps, explanations of how Pathway is used, and any optional agentic integration details.

Domain Focus – Financial Applications (Examples):

Participants may choose any domain (finance, healthcare, e-commerce, etc.) where real-time data makes a difference. For instance, in finance, potential applications include:

- **Compliance:** Automate the interpretation of new regulations (e.g., AML, MIFID) and flag changes via alerts.
- **Due Diligence:** Extract key metrics from pitch decks or risk reports.
- **Analyst Reports:** Generate dynamic investment analyses that reference real-time data.
- **Asset Management:** Merge diverse ESG data into up-to-date compliance summaries.

Optional Add-Ons (Agentic Framework)

- You may integrate an AI agent framework of your choice (e.g., LangGraph, Crew AI, AutoGen, OpenAI Swarm).
- The agent can orchestrate multi-step actions, call external tools, or handle complex conversation flows. (Not a deal-breaker)
- Your real-time Pathway pipeline should still remain the single source of truth for up-to-date contextual data.

Note: We encourage agent experimentation, but the primary focus is on **demonstrating Pathway's real-time RAG capabilities**. You'll get **extra points** for a clever or unique agent use case, but it is *not mandatory* to implement full-blown agent orchestration.

Non-Negotiable: Participants integrating an AI agent framework must deploy the custom agentic workflow by exposing their agent logic via a REST API endpoint using, ensuring seamless interaction with the real-time RAG pipeline powered by Pathway

---

4. Overall Evaluation Framework

I. Technical Implementation (70%)

1. **Real-Time Ingestion & Indexing (35%)**
    - **Key Focus:** How effectively does the solution use Pathway for real-time data ingestion and indexing?
    - **Metrics:**
        - Integration with real-time data sources
        - Low latency in updating the document store
        - Robustness of hybrid retrieval (BM25 + semantic search)
2. **RAG Pipeline Implementation (30%)**
    - **Key Focus:** Quality and correctness of the retrieval-augmented generation pipeline.
    - **Metrics:**
        - Accuracy and relevance of retrieved context
        - Reduction in LLM hallucinations
        - Clear and logical pipeline architecture
3. **Deployment & REST API (15%)**
    - **Key Focus:** Packaging and deployment via Pathway's REST API endpoint.
    - **Metrics:**
        - Ease of setup and replication (Docker, clear instructions)
        - Functionality and clarity of the exposed API
        - Code quality and documentation
4. **UI & Demo Functionality (10%)**
    - **Key Focus:** A basic, functional UI to demonstrate end-to-end usage.
    - **Metrics:**
        - User-friendliness and ease of interaction
        - Ability to observe real-time data changes
        - Integration with the backend RAG pipeline
5. **Optional Bonus – Agentic Integration (Up to +15% Bonus)**
    - **Key Focus:** Creative integration of external agent frameworks (e.g., LangGraph, Crew AI, AutoGen, OpenAI Swarm).
    - **Metrics:**
        - Originality and depth of agent logic
        - Stability and added value from the agent orchestration
        - Clear demonstration of how the agent leverages the Pathway data store

---

II. Presentation (30%)

1. **Slide Content (15%)**
    - **Key Focus:** Clarity and thoroughness of the presentation slides.
    - **Metrics:**
        - **Problem Statement Explanation:** Clearly introduce the chosen problem (e.g., compliance, due diligence) and its relevance.

        - **Tech Stack Overview:** Detail how Pathway was integrated with other components (e.g., DeepSeek/LLM Providers, Ollama, Streamlit).
        - **Innovative Elements:** Highlight key innovations such as agentic logic, real-time alerts, or unique data transformations.
        - **Structure & Visuals:** Slides should be well-organized, visually engaging, and informative.
2. **Demo (15%)**
    - **Key Focus:** Demonstration of the system in action (live or recorded).
    - **Metrics:**
        - **Live/Recorded Demo Quality:** The demo should clearly showcase the system's functionality, referencing relevant parts of the code from the GitHub repo.
        - **Smooth Operation:** Demonstrate real-time updates, API interactions, and UI functionality.
        - **Clarity & Engagement:** The demo should be easy to follow, highlighting the unique aspects of the solution, and effectively tying back to the problem statement.

---

5. Bonus Ideas

- **Automated Alerts:** Trigger Slack/email alerts if a newly ingested document changes the answer to a compliance query.
- **Enhanced Summaries:** Summarize key financial metrics (e.g., EBITDA, ROI) in a structured table.
- **Extending Pathway:** Extend an existing class in Pathway to unlock new capabilities.
- **Handling Complex Data:** Show how to process non-textual formats (tables, charts) using vLMs where pure text-based LLMs might struggle.
- **Agentic Error Handling:** Implement fallback mechanisms (e.g., switching to a backup data source) if a primary feed fails.
- **Cutting-Edge Implementations:** Explore state-of-the-art models (e.g., Gemini 2.0) or design a "Faraday cage" approach with zero external API dependency.

6. All the resources you need to get started:

- RAG Introductory Blog
- Building your first Realtime RAG pipeline with Pathway:
    - Building with OpenAI
    - Building with Gemini/Other LLMs
    - Implementation of Pathway with LlamaIndex
    - Implementation of Pathway with LangChain
- Keywords to target:
    1) Real-Time Data for Generative AI
    2) RAG with data streaming
    3) Event driven Agent/RAG orchestration
- Building a Realtime Agentic RAG pipeline using LangGraph and Pathway

- LLM Tooling (Pathway's core software development kit for building a custom RAG pipeline, integrating Pathway into your existing codebase, or doing deep customizations)
- API Documentation for Pathway LLM xpack
- Pathway Developer Documentation: Link to Pathway Developer Docs
- How to deploy agents with Pathway?
    - Here you will see how you can build custom endpoints using Pathway RAG classes. There are two ways to serve agents: using the serve_callable API (which is easier to manage and recommended) or with an external web server like FastAPI. If you prefer, you can start with an external web server and move the endpoint to Pathway later.
- Tips for resolving doubts?
    - Leverage Gen AI wisely. If you see difficult-to-comprehend error messages, the least you should do is ask the query on Bard / Bing AI search, etc.
    - Utilize the #get-help channel on Pathway's Discord, if needed. However, given the nature of the competition, we wouldn't be able to share direct answers.