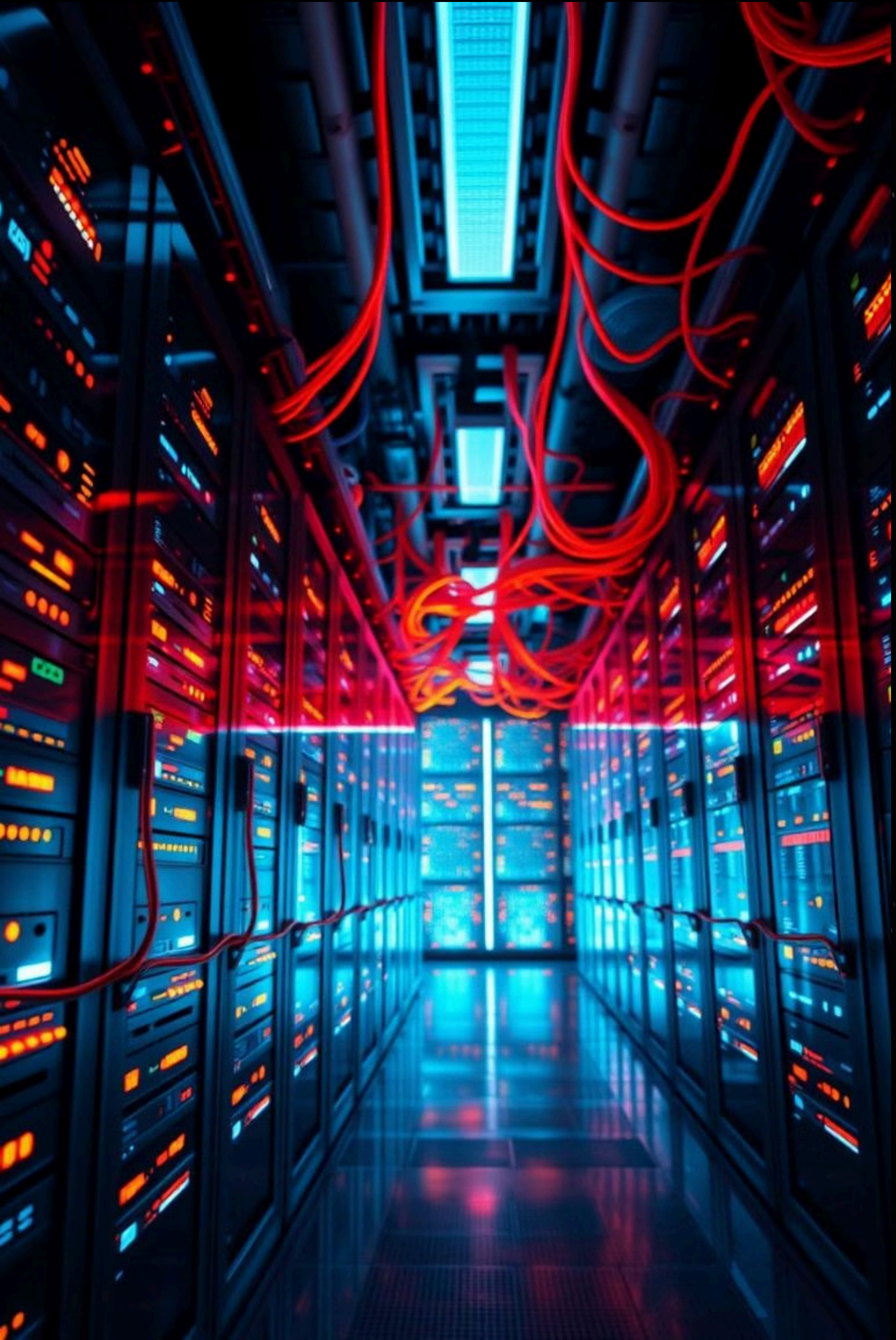

Creating a Real-Time RAG Application with Pathway

This presentation explores how to build a real-time Retrieval Augmented Generation (RAG) application using Pathway. We will delve into leveraging Pathway's vector store, agentic frameworks, and real-time data updates to create an end-to-end RAG system. Our objective is to demonstrate how Pathway can orchestrate data ingestion, incremental indexing, and REST API deployment for a truly dynamic and responsive system. The presentation will also focus on showcasing how real-time pipelines adapt to data updates, ensuring the freshest context for every query.

S by Shiveshwar Sah

Introduction & Key Challenges

Advanced RAG System

- Integrates Pathway's Vector Store
- Collaborates with multi-agent frameworks
- Analyzes data from textual sources in real-time

Core Requirements

- **Query Understanding:** Decomposition, clarification, disambiguation
- **Retrieval Planning:** Balances complexity vs. efficiency
- **Retrieval Mechanisms:** Advanced semantic & hierarchical mapping
- **Verification & Correction:** Prunes irrelevant info, resolves inconsistencies, flags harmful content
- **Multi-Hop Query Analysis:** Iterative retrieval/tool usage

Financial Domain Focus

1

Agentic RAG Pipeline

Tailored for **financial data** (textual content, numerical tables). Ingests multi-modal data for holistic analysis.

2

Complexity & Criticality

Precision and accuracy are paramount in this high-stakes decision-making environment.

3

Key Value Proposition

Minimizes risks by delivering **reliable** & **clear** outputs. Acts as an **AI-powered financial analyst** for leadership decisions.

System Architecture

1

PDF Parsing

Extracts text, tables, images. Converts tables into HTML format for easy processing.

2

Recursive Chunking

Splits long text into manageable segments, optimizing retrieval efficiency.

3

Pathway's Vector Store

Stores chunked data. Enables efficient real-time querying for relevant information.

4

Agentic Retrieval

Agents query relevant chunks for answers, improving accuracy and reducing noise.

Agent Decision-Making & Tool Selection

LLM Evaluation

Assesses retrieved chunks for **relevancy**. Decides if more context is required for accurate responses.

Tool Selection

Options include: **Yahoo Finance** (market data), **Python Calculator** (computational tasks), **Edgar Tool** (SEC filings), **Bing Web Search** (general web info).

Iterative Retrieval

If the LLM identifies gaps, it calls appropriate tools. Minimizes irrelevant or incomplete answers.

Real-Time Endpoint & Scaling

Deployment

Pathway's serve method hosts the RAG end-to-end endpoint, providing a streamlined approach for handling user queries in real time.

Scalability

Designed to manage **high demand** scenarios.

Automatically adapts to **data updates** in the vector store.

Fresh Context

Ensures current and updated information is used for every query, improving accuracy and relevance.