DATA+AI SUMMIT
BY databricks

DESIGNING ENTERPRISE RAG SYSTEMS

Erik Widman, Ph.D. – Lead ML Director @ CVS Health
6.10.24

CVS HEALTH AT A GLANCE

From humble beginnings to F6 healthcare company

HISTORY

- 1963 – Founded in Lowell, MA
- 2006 - Expansion into healthcare services
- 2018 – Acquisition of Aetna
- 2020s – Technology and value based healthcare

BUSINESS

- F6 company
- Revenue $322B
    - 11% YoY increase
- Profit $8.3B
    - 100% YoY increase

SCALE

- 300,000+ corporate colleagues
- 200,000+ retail employees
- ~10,000 retail stores

OUR LINES OF BUSINESS

CVS Health is much more than a pharmacy...

CVS pharmacy

Retail Pharmacy

- Prescription and over-the-counter drugs

CVS caremark

Pharmacy Benefits Management

- Manage prescription drug benefits for company employees

minute clinic

Healthcare Services

- Walk-in medical services

CVS specialty

Specialty Pharmacy

- Advanced medications for chronic diseases

aetna

Health Insurance

- Medical, dental, vision, prescription drugs

Oak St. Health

signifyhealth

In-Home and Primary care

- Value-based care

FINDING INFORMATION @ CVS HEALTH

Challenges with being a large company

Many knowledge sources

- SharePoint
- ServiceNow
- Confluence
- Homegrown solutions

New vs. existing colleagues

- Don’t know where to search
- Existing search performs poorly
- Domain expertise + *Word of mouth*

Speed of change

- Rebranding + reorganization
- Continuous updates of policies and documentation
- Document migration

SIMPLIFY
+
UNIFY

RETRIEVAL AUGMENTED GENERATION

Limitations of POC RAG systems

- Static Data Sources
- Uniform Content from a single source
- ”Small” datasets
- Scaling limitations for a small number of users

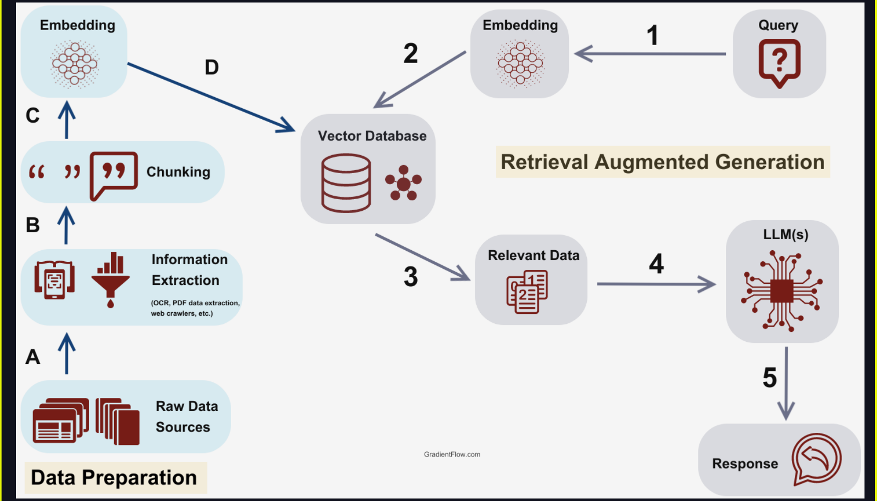


Image source: Gradientflow.com

CHALLENGES WITH SCALING RAG

Knowledge management at scale is difficult

Document Volumes

- Tens of millions of documents at a large company
- Difficulty measuring # of docs @ scale

Multiple Data Sources

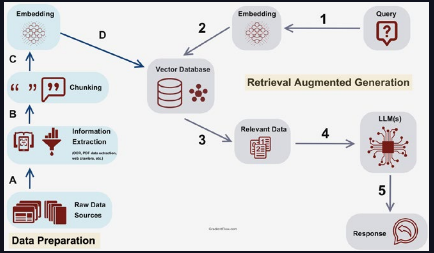
- What is a knowledge source?
- Mapping knowledge sources
- Where do you start?

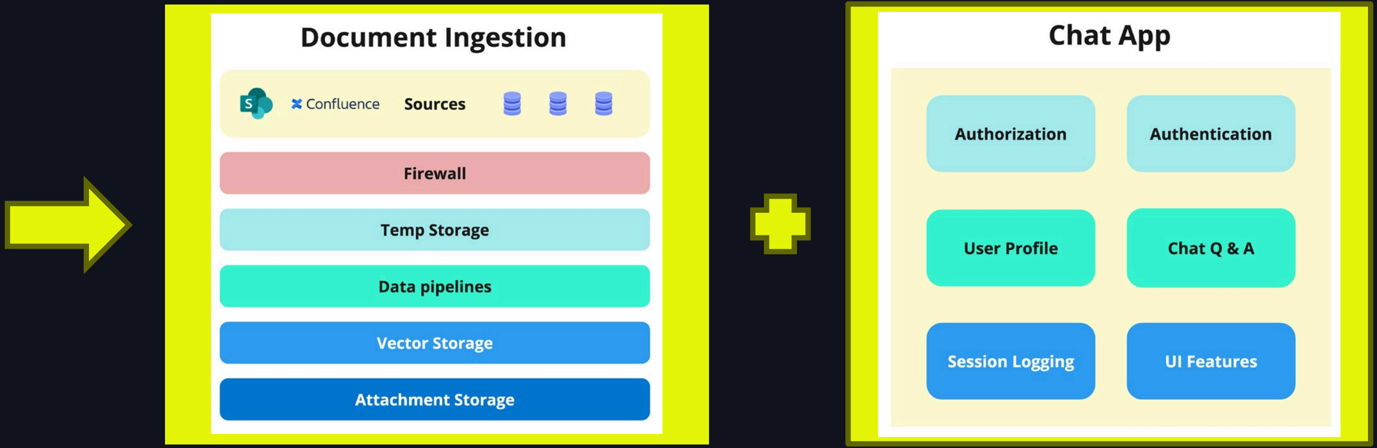
Document lifecycles

- Every document has a lifecycle
- What documents are useful
- Public vs private

RETHINKING RAG AS A PRODUCT

The first step to scaling is a long-term product mindset




Document Ingestion

Confluence Sources

Firewall

Temp Storage

Data pipelines

Vector Storage

Attachment Storage

Chat App

Authorization

Authentication

User Profile

Chat Q & A

Session Logging

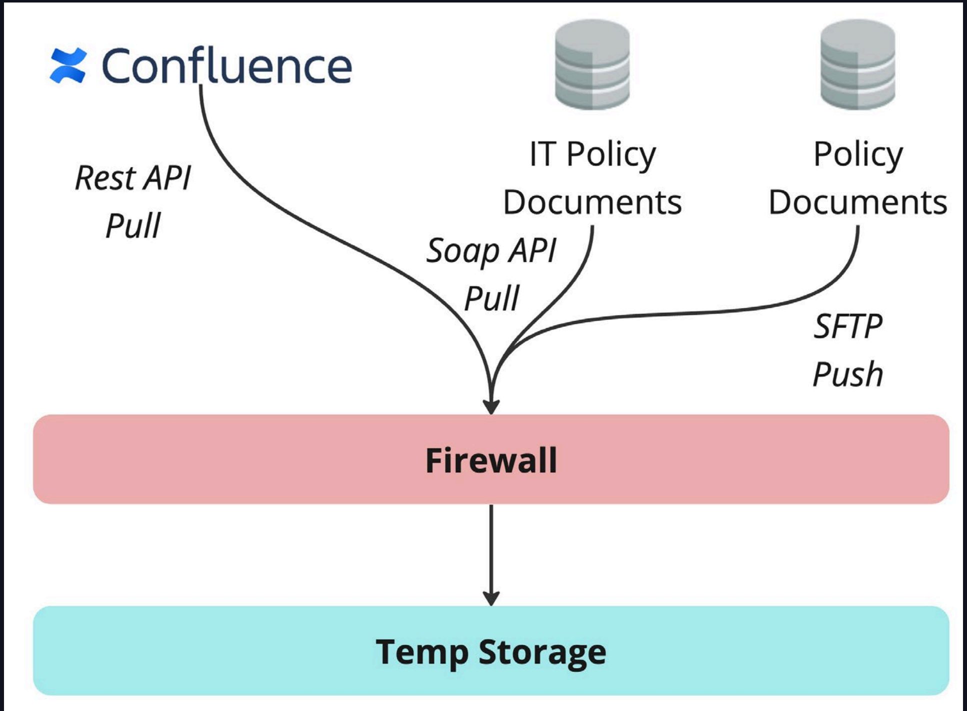
UI Features

- OKRs + analytics for each product
- UI points of entry for maximum adoption

DOCUMENT INGESTION

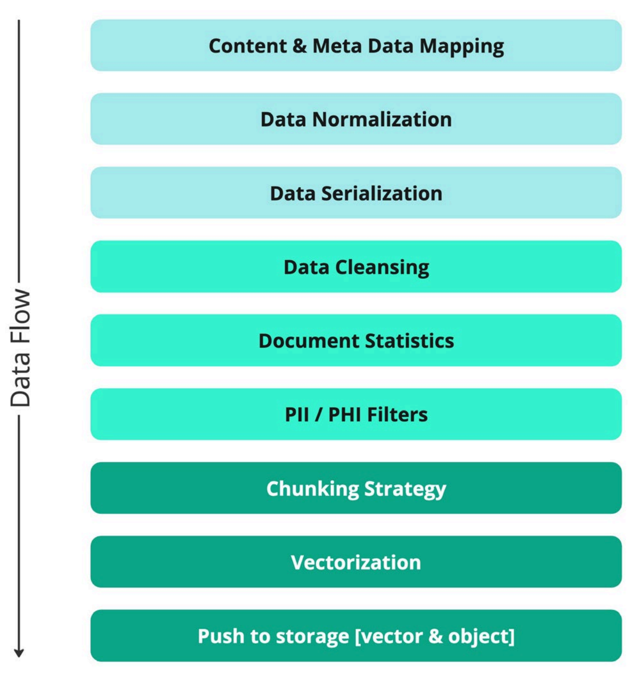
There is no “one-size-fits-all” data connector

- Need to connect directly to sources
- Data ingestion mechanisms



DATA PIPELINES

Normalize your data from various document types

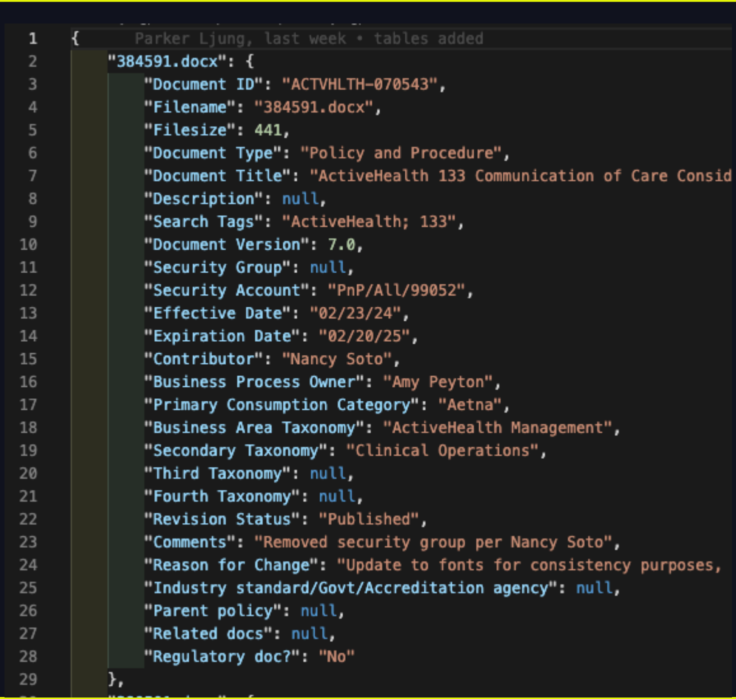


- **Custom pipelines required for each source**
- **Data normalization** – standardize data from various sources and file types
- **Serialization** – JSON convenient to work with
- Data cleansing
- Chunking strategies

METADATA SCHEMA

Normalize the metadata schema across sources

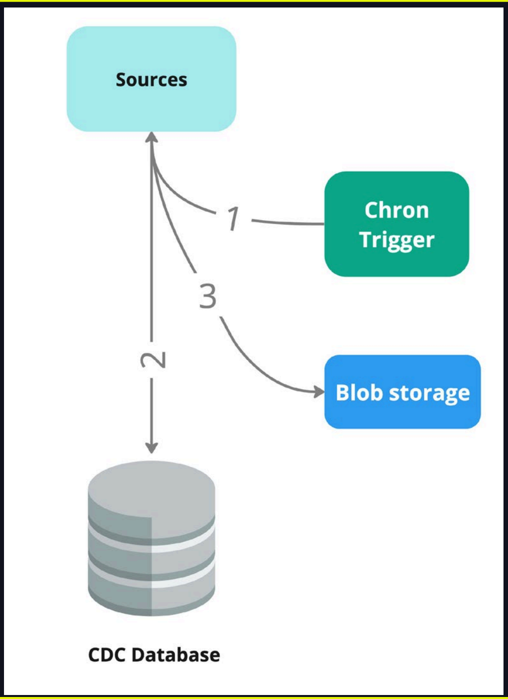
- Available source metadata schema vs. normalized metadata schema
- Metadata used for:
    - Filtering
    - CDC trigger
    - Access control
    - Related content
    - Document information



```
{       Parker Ljung, last week • tables added
    "384591.docx": {
        "Document ID": "ACTVHLTH-070543",
        "Filename": "384591.docx",
        "Filesize": 441,
        "Document Type": "Policy and Procedure",
        "Document Title": "ActiveHealth 133 Communication of Care Consid
        "Description": null,
        "Search Tags": "ActiveHealth; 133",
        "Document Version": 7.0,
        "Security Group": null,
        "Security Account": "PnP/All/99052",
        "Effective Date": "02/23/24",
        "Expiration Date": "02/20/25",
        "Contributor": "Nancy Soto",
        "Business Process Owner": "Amy Peyton",
        "Primary Consumption Category": "Aetna",
        "Business Area Taxonomy": "ActiveHealth Management",
        "Secondary Taxonomy": "Clinical Operations",
        "Third Taxonomy": null,
        "Fourth Taxonomy": null,
        "Revision Status": "Published",
        "Comments": "Removed security group per Nancy Soto",
        "Reason for Change": "Update to fonts for consistency purposes,
        "Industry standard/Govt/Accreditation agency": null,
        "Parent policy": null,
        "Related docs": null,
        "Regulatory doc?": "No"
    },
```

CONTINUOUS DATA INTEGRATION

Documents are dynamic



- CDC Mechanisms for various sources
- Data ingestion rate strategies
  - "Real-time"
  - Batch
- New documents
- Edited document
- Deleted documents

MICROSERVICES

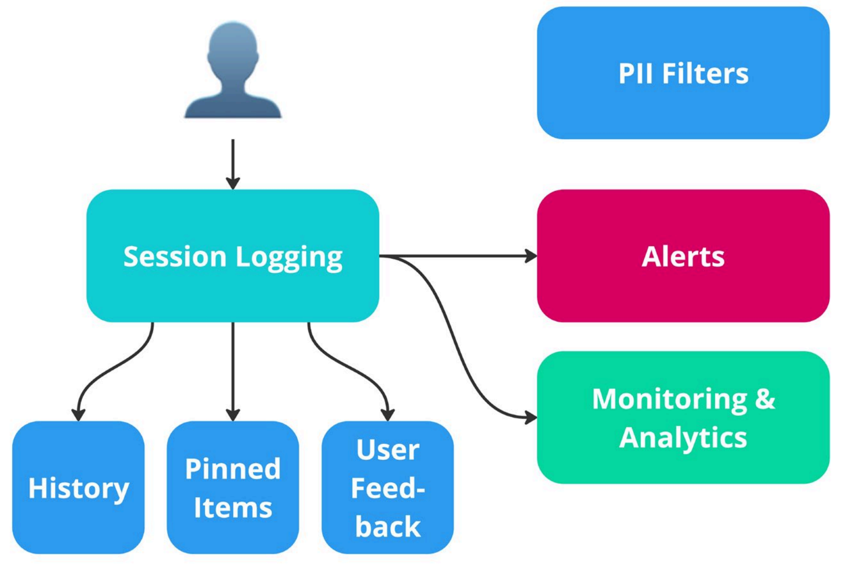
We need containers to handle dynamic workloads

- Document ingestion
- Chat app
- Kubernetes
- GitHub actions
- Helm

SAFETY AND MONITORING

Keep users safe and flag for inappropriate behavior

- Session logging
- Alerts
- Analytics
- PII/PHI filters
- User feedback



OPTIMIZATION STRATEGIES

Great RAG results requires a combination of strategies

Pre-retrieval strategies	Retrieval strategies	Post-retrieval strategies
• Improving the quality of your indexed data • Chunk optimization • Query rewriting	• Using alternative search methods • Using different embedding models • Small2big, recursive, or context-aware retrieval • Hierarchical retrieval	• Reranking or scoring retrieved chunks • Information compression

RAG EVALUATION FRAMEWORKS

Quantitative vs human evaluation

Manual evaluation

- Tests data set
- Answer relevance to query
- Context relevance
- Iterate

SME Review

- Domain expertise
- Identify relevant documents for sources

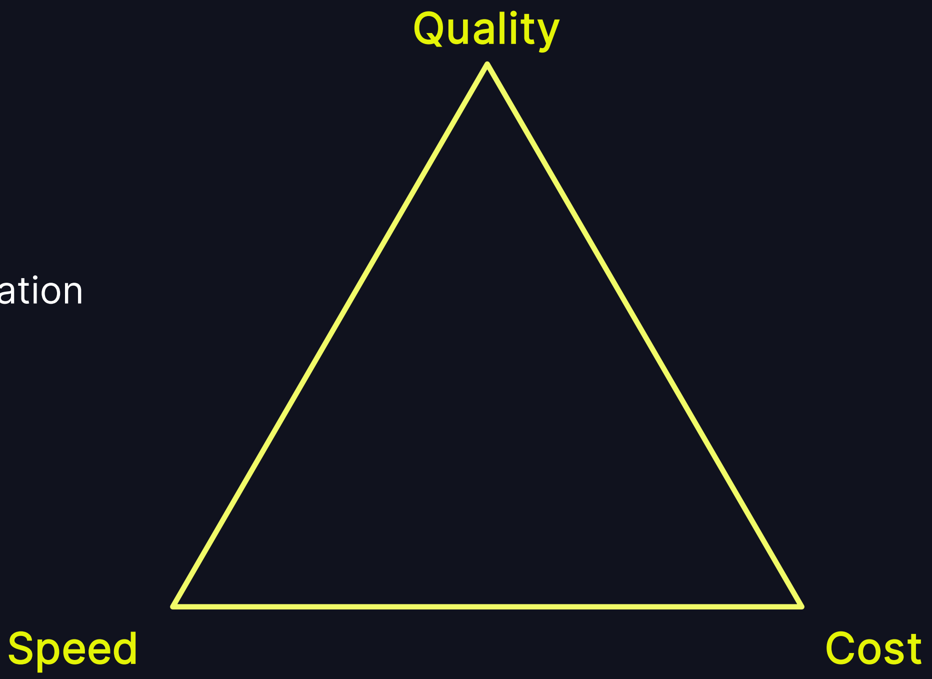
LLM Review

- Truelens
- Quantitatively score response with LLM

THE LLM HOLY TRINITY

Finding the sweet spot

- Using smaller, faster models for some steps
- Making intermediate steps run parallel
- Have the LLM make choices instead of generation
- Implementing caching



HOW TO GET STARTED

Knowledge discovery is a product, not a technology

Discovery

- Source identification
- Source Owners
- Types of knowledge
- Connectivity
- ROI

Don’t boil the ocean

- Start with 2-3 sources
- Wide vs deep
- Optimization vs expansion

Build to scale

- Build a platform
- Legos
- Initial build vs steady state

WHAT IS TRUTH?

DATA+AI SUMMIT