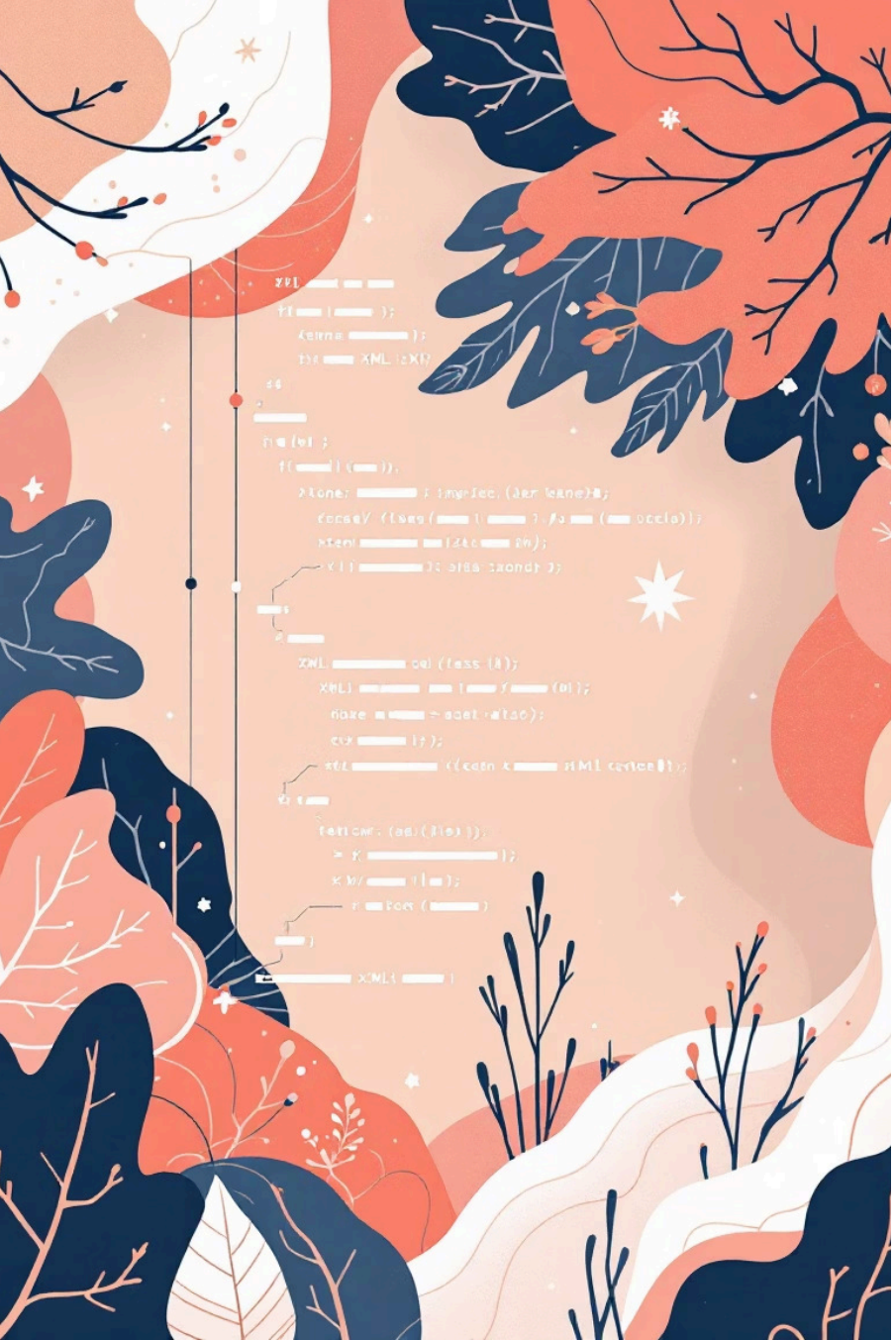
XML to CSV with RAG-based Insights

Transforming Legacy XML Data into Intelligent Analytics

Tech Stack: Python, Pandas, LangChain, FAISS, OpenAI, XML Parser



Problem & Context

The Challenge

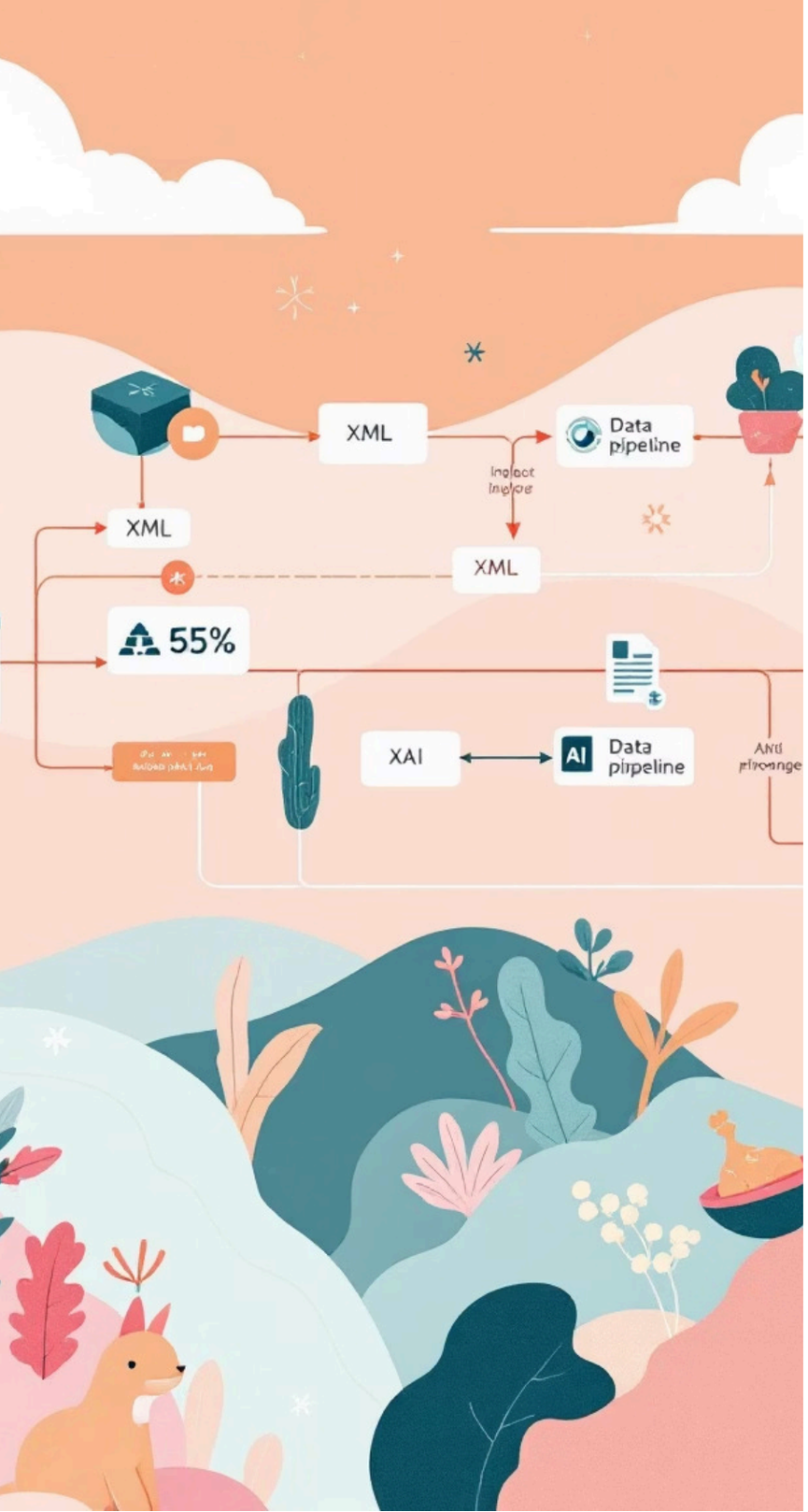
XML data presents significant obstacles for modern analytics teams. Its hierarchical structure, inconsistent formatting, and nested elements make direct querying incredibly complex and time-consuming.

Our Solution

Transform XML into structured CSV format, then implement Retrieval-Augmented Generation (RAG) to enable natural language querying for instant insights.

Business Impact

Particularly valuable for organisations managing legacy XML systems, enabling analysts to extract insights through conversational queries rather than complex parsing.



Approach & Architecture

Our streamlined approach transforms complex XML structures into intelligent, queryable systems through a systematic pipeline.

Parse XML

Extract and normalise XML data structures

Convert to CSV

Transform into structured tabular format

Text Chunking

Break data into optimised retrieval segments

RAG Query

Enable natural language insights

Key Innovation: Metadata preservation ensures complete traceability - column names and XML tags maintain data lineage throughout the transformation process.

Implementation Details

Core Workflow

01

XML Parsing & CSV Creation

Parse XML structure and export as structured CSV maintaining data relationships

02

Text Structuring

Convert CSV rows into formatted text using [column_name] value pattern for clarity

03

Intelligent Chunking

Split text into 300-character chunks with 50-character overlap for optimal retrieval

04

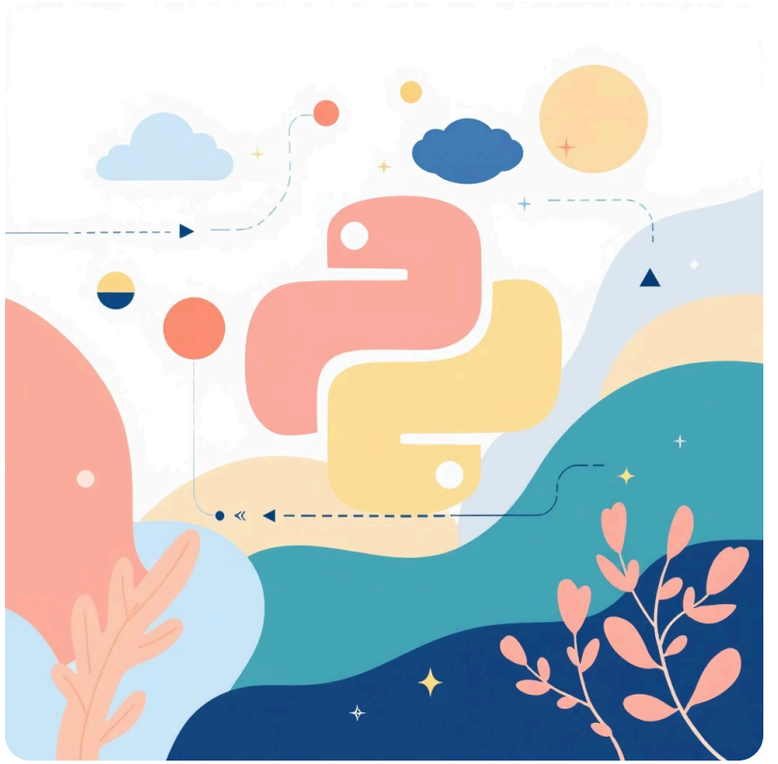
Vector Storage

Generate embeddings and store in FAISS for high-performance similarity search

05

Query Processing

Use ChatOpenAI with prompt engineering to enforce source citation and accuracy



Prompt Engineering: All responses must include source information, ensuring complete transparency and traceability in analytics results.

Results & Benefits

Successful Transformation

Complex XML datasets now accessible through intuitive natural language queries, eliminating technical barriers for business analysts and stakeholders.

Enhanced Traceability

RAG pipeline maintains complete data lineage, with every insight traceable back to its original XML source for audit compliance and verification.

Intelligent Accuracy

Advanced prompt engineering ensures responses are both accurate and transparent, with mandatory source citations building user confidence in results.

> "Transform your legacy XML data into a conversational analytics platform that your entire team can use effectively."

Future Enhancements

Adaptive Chunking

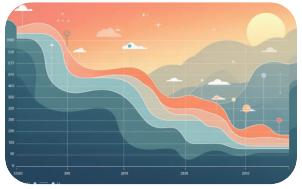
Implement semantic chunking strategies that adjust to variable data lengths and content types for improved retrieval accuracy.

Graph RAG Integration

Enable cross-CSV relational queries through graph-based RAG, connecting disparate datasets for comprehensive business insights.

Enhanced Metadata

Preserve hierarchical XML tag structures in CSV mappings, maintaining complete data context and relationships.



Better Vector Store

We can use vector databases like Pinecone when dealing with industry grade implementation of Vector DB's