esprit Se former autrement	*Élaboré par :*	*Equipe PI-DS*
	Année Universitaire :	*2024-2025*

N.B: ***Ce document est susceptible d'être ajusté et d'accueillir de nouvelles idées et fonctionnalités que vous pourriez proposer, en vue d'une éventuelle expansion de son contenu.***

Technical Architecture Document (TAD) : Due Diligence on Cryptocurrency and Digital Assets Funds

Project Overview

This project aims to:

- Collect data from multiple sources on digital assets and cryptocurrencies.
- Build structured pipelines for data engineering.
- Generate a dynamic question bank using reverse engineering and prompting techniques with Generative AI (GenAI).
- Develop a system to answer questions based on multiple documents from crypto funds using GenAI.
- Generate detailed reports in PowerPoint format based on the data and insights.

Example of the project in the market :

https://www.youtube.com/@Autogen_AI

https://www.youtube.com/watch?v=w2lelkUdXJw

Bibliography :

https://github.com/ico-check/ico-check

https://theholycoins.com/blog/how-to-do-due-diligence-in-cryptocurrency-investments

https://blogs.cisco.com/security/cryptocurrency-and-blockchain-security-due-diligence-a-guide-to-hedge-risk

https://steemit.com/bitcoin/@mtkander/ico-analysis-framework

https://crypto.com/glossary/fr/customer-due-diligence-cdd

https://www.investopedia.com/financial-responsibility-and-crypto-due-diligence-8385090

https://www.sanctions.io/blog/crypto-due-diligence

https://www.ey.com/en_us/insights/financial-services/token-due-diligence-a-structured-approach-to-digital-asset-risk

https://www.centrl.ai/resources/operational-due-diligence-for-crypto-investing/

https://www.scorechain.com/resources/crypto-glossary/customer-due-diligence

https://www.integrityriskintl.com/services/cryptocheck/

https://www.ftitechnology.com/solutions/decentralized-due-diligence

https://www.compilot.ai/academy/glossary/enhanced-due-diligence-edd-a-comprehensive-guide-for-crypto-compliance

https://bitaml.com/2019/02/18/edd-crypto-msbs/

https://cryptoinvestigators.com/crypto-investigations-due-diligence/

https://www.fticonsulting.com/insights/articles/due-diligence-cryptocurrency-strategy-investing-rebounding

https://www.elliptic.co/platform/discovery

List of Funds :

- Pantera Capital: ~$4.2 billion, Hedge Fund & Venture Capital, an early leader in crypto investing across Bitcoin, blockchain ventures, and tokens

- Polychain Capital: ~$6.6 billion, Hedge Fund, specializes in cryptocurrency protocols and blockchain startups.

• Brevan Howard Digital: ~$2.3 billion, Hedge Fund, focuses on digital assets as a division of Brevan Howard with strong performance.

• Nickel Digital Asset Management: ~$200 million, Hedge Fund, London-based firm delivering consistent crypto investment returns.

• Fasanara Digital: ~$150 million, Hedge Fund, invests actively in digital assets as part of Fasanara Capital.

Their Strategies : long / short / arbitrage / techno / infra… etc etc These are completely different to ETF (they are not subject to regulations…)

1. System Components

1.1 Data Collection Layer

Objective: Collect data related to digital assets and cryptocurrencies from diverse sources.
Components:

- **API Integrations:**
  - CoinMarketCap, CoinGecko, Glassnode APIs for market and on-chain data.
  - Blockchain Node APIs (e.g., Etherscan) for transaction and wallet data.
- **Web Scraping:**
  - Tools: BeautifulSoup, Selenium.
  - Targets: Crypto fund websites, regulatory documents, and news sources.
- **Blockchain Nodes:**
  - Ethereum/Bitcoin nodes for direct blockchain data.
- **Data Storage:**
  - **Relational Database:** PostgreSQL for structured data.
  - **NoSQL Database:** MongoDB for unstructured data.

1.2 Data Engineering Pipeline

Objective: Create robust pipelines for data preprocessing and transformation.
Components:

- **Data Transformation:**
  - Tools: pandas, GenAI.
  - Steps: Data cleaning, normalization, type conversions, deduplication.
- **Data Aggregation & Enrichment:**
  - Aggregate trends, benchmarks, and enriched datasets.
- **Data Validation:**
  - Ensure consistency across sources.

1.3 Question Bank Generation

Objective: Reverse engineer documents to generate a dynamic question bank using GenAI.
Components:

- **Document Parsing:**
  - Tools: pdfplumber, PyMuPDF, Apache Tika.
  - OCR (if necessary): Tesseract for image-based text extraction.
- **LLM Prompting:**
  - Model: GPT-3 or fine-tuned variants.
  - Templates: Generate questions on investment strategies, risks, compliance, etc.
- **Storage:**
  - Database for structured questions categorized by type and topic.
- **Question Ranking:**
  - Tools: BERT, RoBERTa to rank questions based on relevance.

1.4 GenAI-Powered Q&A System

Objective: Build a system to answer questions based on multiple documents using GenAI.
Components:

- **Text Embedding:**
  - Models: Sentence-BERT, GPT-3 embeddings.
  - Storage: FAISS, Elasticsearch for vector search.

- **Answer Generation:**
  - Fine-tuned LLMs for domain-specific responses.
  - Retrieval-augmented generation (RAG) for context-aware answers.
- **Backend Integration:**
  - Frameworks: FastAPI, Django DRF

1.5 Report Generation

Objective: Generate comprehensive PowerPoint reports dynamically.
Components:

- **Data-Driven Content:**
  - Extract key insights (e.g., fund performance, compliance metrics).
  - Visualize data with matplotlib, Plotly, Seaborn.
- **Slide Generator:**
  - Tools: python-pptx.
  - Templates for consistent formatting (e.g., Title Slide, Performance Overview).
- **Export Functionality:**
  - Allow users to download the final report as a .pptx file.

1.6 User Interface

Objective: Provide an intuitive interface for interacting with the system.
Components:

- **Frontend:**
  - Frameworks: Dash-plotly
  - Features: Upload documents, query interface, report download option.
- **Feedback Mechanism:**
  - Allow user feedback to improve system accuracy.

2. System Architecture

Layered Diagram:

1. **Data Collection Layer:** APIs, Web Scrapers, Blockchain Nodes.
2. **Data Engineering Pipeline:** ETL, Transformation, Validation.
3. **GenAI Question Bank:** Parsing, Prompting, Ranking.
4. **Q&A System:** Text Embedding, Retrieval, Answer Generation.
5. **Report Generator:** Data Insights, Visualizations, PPTX Export.
6. **User Interaction Layer:** Web UI for seamless user experience.

3. Technology Stack

Backend: Python, FastAPI, Django RestFramework
Frontend: Powerpoint presentation pptx, Dash Plotly
Data Storage: PostgreSQL & PgVector, MongoDB, FAISS, Elasticsearch.
AI/ML Models: GPT-3, BERT, Sentence-BERT.
Visualization: matplotlib, Plotly.
Report Generation: python-pptx.
Orchestration: Docker, Kubernetes.

4. Deployment and Scalability

- **Cloud Platform:**
    - Use managed services for databases, storage, and AI models.
- **Containerization:**
    - Docker for packaging microservices.
- **Orchestration:**
    - Kubernetes for scaling microservices.
- **CI/CD:**
    - Implement pipelines using GitHub Actions, Jenkins, or GitLab CI/CD.