



SSI AI MASTERPLAN

AI-First Intelligence





MỤC LỤC

TẦM NHÌN VỀ AI

02

ỨNG DỤNG AI TẠI SSI

03

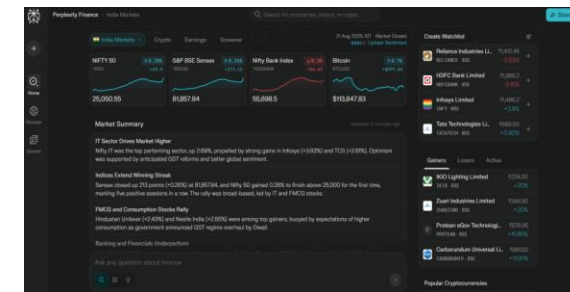
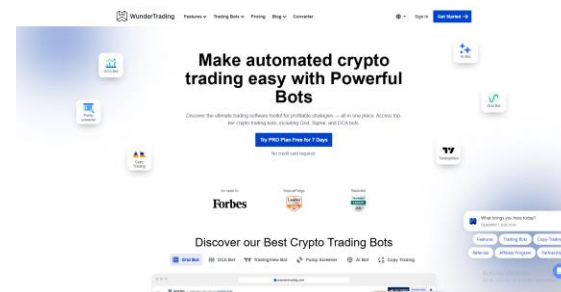
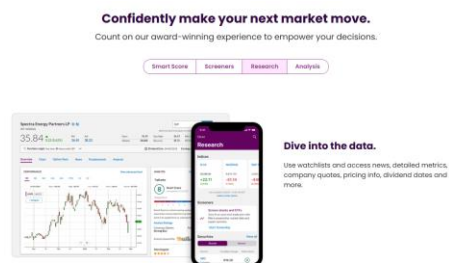
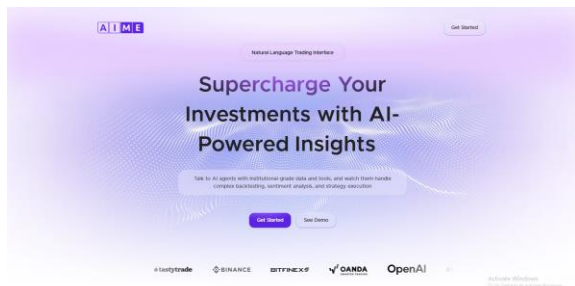
GIẢI PHÁP & LỘ TRÌNH TRIỂN KHAI





TÀM NHÌN VỀ AI

AI không phải là tương lai, AI là hiện tại, AI đang liên tục phát triển hơn mỗi ngày



AI tạo ra cơ hội cho các công ty nhỏ hơn bắt kịp và trở thành thách thức của các doanh nghiệp lớn

80%

Giao dịch tài chính toàn cầu sẽ được thực hiện qua các tác nhân AI (AI Agent) tới năm 2030

Khách hàng đang khao khát AI, Và đó là nhu cầu mới của thị trường

80%

Khách hàng sẵn sàng chuyển đổi từ một tổ chức tài chính đã gắn bó nếu như tìm thấy một giải pháp AI

Accenture, 2024

70%

Người được hỏi muốn ngân hàng và các tổ chức tài chính sử dụng dữ liệu giao dịch một cách hiệu quả hơn thông qua các phân tích AI

Personetics, 2024

25%

Lượng truy vấn qua nền tảng chatbot của các ngân hàng đã tăng lên trong 1 năm qua

Bank of America, 2025

Quan trọng hơn cả việc chúng ta cạnh tranh với ai, chúng ta có đang thực sự giúp các khách hàng nâng cao trải nghiệm & hiệu quả giao dịch?

Nhà đầu tư mới

“Tôi cần mọi thông tin dễ hiểu, được đi từ đầu đến cuối từng bước”

Nhà đầu cơ

“Tôi cần thông tin nhanh, gọn, vào đúng thời điểm”

Nhà đầu tư giá trị

“Tôi cần một kịch bản đầy đủ, nhưng không chuyên sâu và bám sát mục tiêu tài chính cá nhân”

Nhà đầu tư kinh nghiệm

“Tôi cần thêm những công cụ ưu việt, tất cả trong một”

Khách hàng cao cấp

“Tôi cần những trải nghiệm cao cấp, toàn diện, tận dụng tối đa sức mạnh của con người & công nghệ”

Môi giới

“Tôi không thể hỏi đáp mọi câu hỏi của các nhà đầu tư. Tôi muốn tập trung vào việc giúp nhà đầu tư đưa ra hành động đầu tư sinh lời”

Vận hành

“Có quá nhiều lỗi phát sinh do con người tạo ra”

Phát triển sản phẩm

“Nên chọn phiên bản thiết kế nào đây?”

**AI không thay thế con người,
AI làm những việc con người
làm chưa tốt, theo một cách
năng suất hơn**

Dễ tiếp cận

Phục vụ được đa dạng các chân
dung và nhu cầu tài chính

Luôn thích ứng

Sẵn sàng học hỏi và điều chỉnh
theo các biến động hay xu hướng

Luôn ở bên

Kết nối và lắng nghe người dùng
mọi nơi, mọi lúc

Hiệu quả

Nhanh, hạn chế rủi ro, hướng tới
sự chính xác và mục tiêu đề ra





ỨNG DỤNG AI TẠI SSI

KHÁCH HÀNG

- Tối ưu trải nghiệm
- Hỗ trợ thông tin
- Tư vấn cá nhân hoá



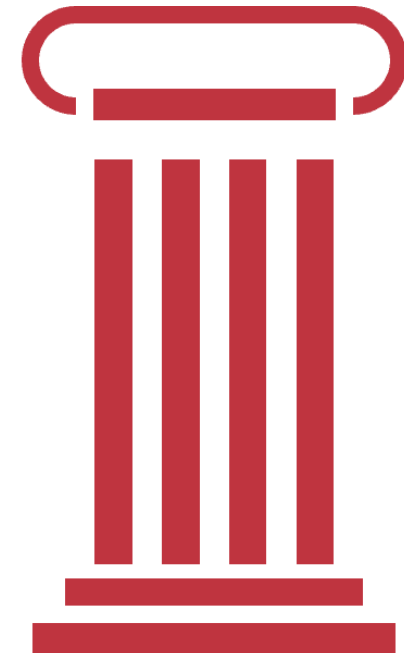
KINH DOANH

- Phân tích & dự báo hành vi
- Thấu hiểu khách hàng qua dữ liệu



VẬN HÀNH


- Tối ưu vận hành
- Trợ lý thông tin nội bộ
- Phát hiện & xử lý sự cố





Customer AI

 Chatbot tư vấn giao dịch, margin, thuế


 Cá nhân hóa danh mục - “Netflix cho đầu tư”


 Research: tóm tắt báo cáo, highlight cổ phiếu


 Hội thoại đa kênh: email, call center, social



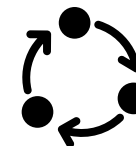
Business & Sales AI

 Lead Scoring: tìm kiếm khách hàng tiềm năng


 Marketing: gợi ý sản phẩm cho từng nhóm khách hàng

 Dự báo hành vi: margin, sản phẩm phái sinh


 Cross/Up-sell: gợi ý thêm sản phẩm



Operational AI

 HR Mind AI: tự động onboard, phân tích hiệu suất, đánh giá – dự báo rời tổ chức

 Security Agent: phát hiện sự cố an toàn thông tin

 Broker Agent: hỗ trợ môi giới về thuế/phí, quy định, hiệu suất realtime

Tên sản phẩm	Chức năng	Tiến độ	Dữ liệu nhạy cảm
Cú Doha phase 1	Hỏi đáp sản phẩm - dịch vụ tại SSI	DONE	NO
Cú Doha phase 2	Cung cấp thông tin thị trường, khuyến nghị từ chuyên gia,	IN PROGRESS	NO
Cú Doha phase 3	Tư vấn đầu tư theo danh mục khách hàng	TO DO	YES
iResearch	Dashboard thông tin cổ phiếu theo PTKT	IN PROGRESS	NO

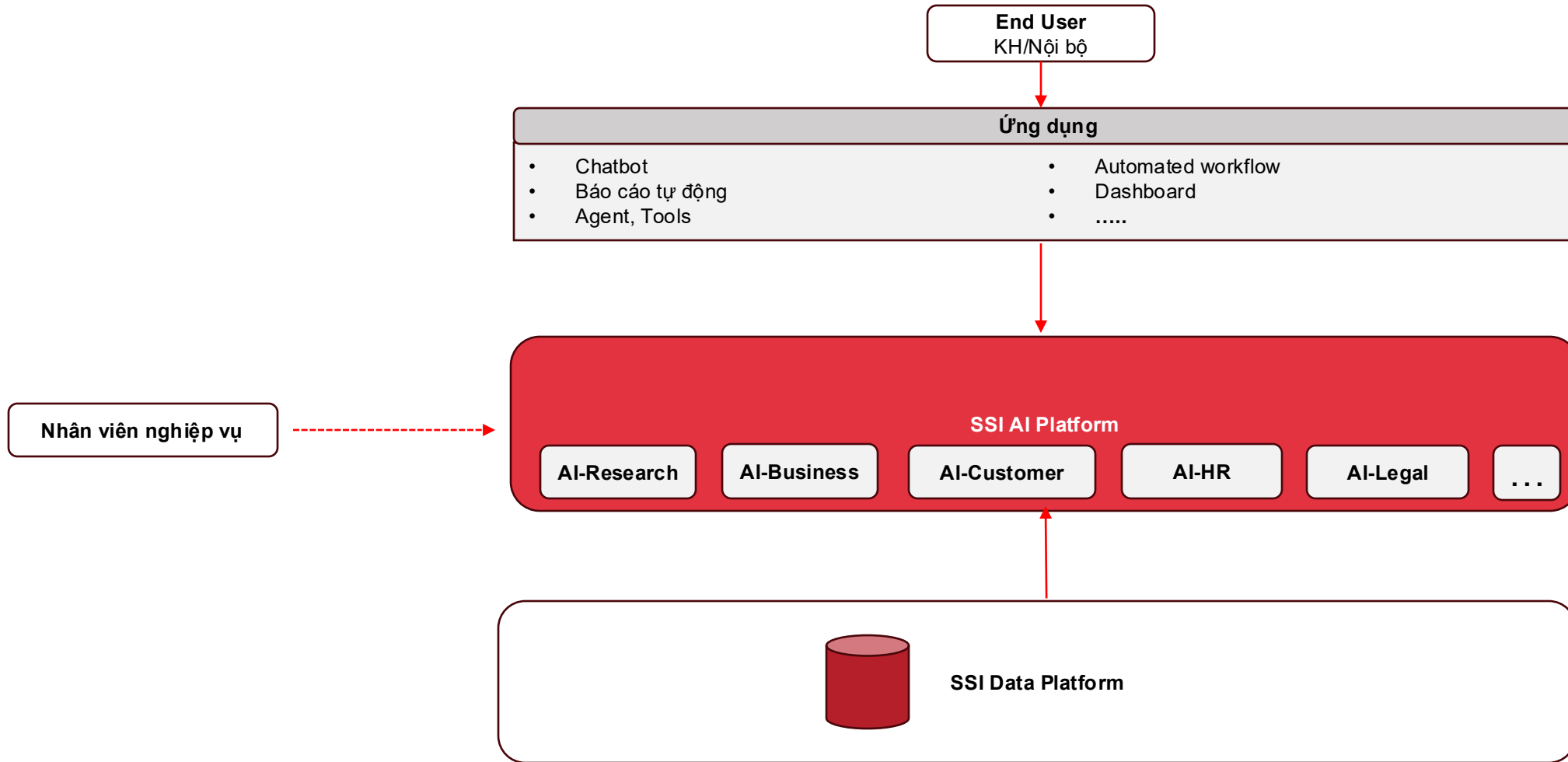
Cần hạ tầng nội bộ để SSI làm chủ hoàn toàn công nghệ, tuân thủ nghị định 13 về dữ liệu khách hàng

Tên sản phẩm	Chức năng	Tiến độ	Dữ liệu nội bộ
Research Assistant	Công cụ hỗ trợ Research viết báo cáo phân tích	S-Shine IDEA	YES
HR Mind AI	Tự động hoá quy trình onboard, phân tích hiệu suất làm việc, đánh giá và ghi nhận thành tích cũng như dự báo nguy cơ rời tổ chức		YES
Security Agent	Phát hiện sự cố an toàn thông tin		YES
Data Agent	Phân tích dữ liệu kinh doanh, phân khúc khách hàng, tìm kiếm khách hàng tiềm năng và dự báo rời bỏ		YES
Broker Agent	Hỗ trợ môi giới giải đáp thuế/phí, quy định, nghiệp vụ giao dịch, hiệu suất môi giới theo thời gian thực..		YES

Nhu cầu lớn, cần hạ tầng nội bộ để bảo mật dữ liệu, tối ưu chi phí phát triển và hiệu suất

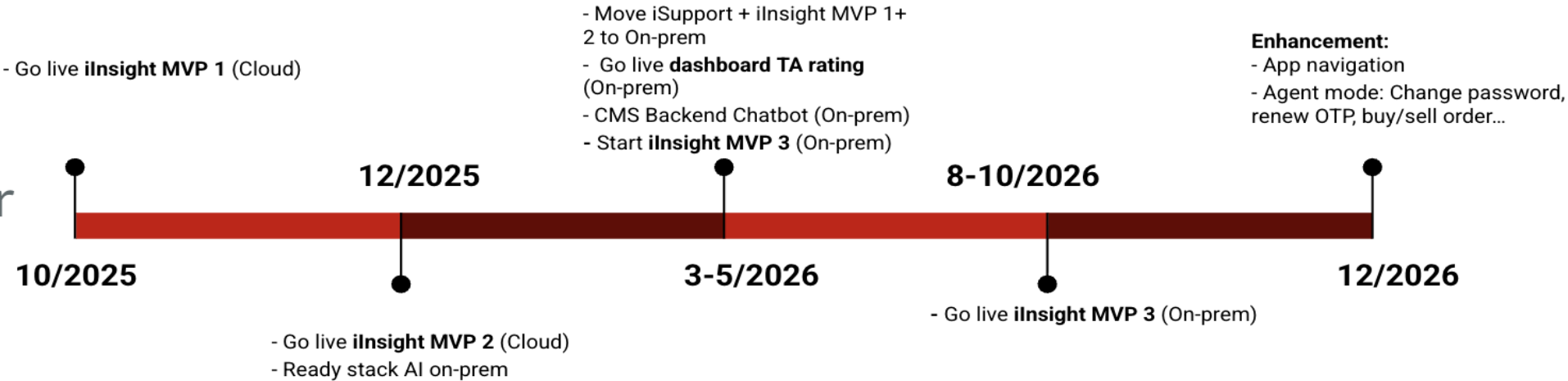


GIẢI PHÁP & LỘ TRÌNH TRIỂN KHAI

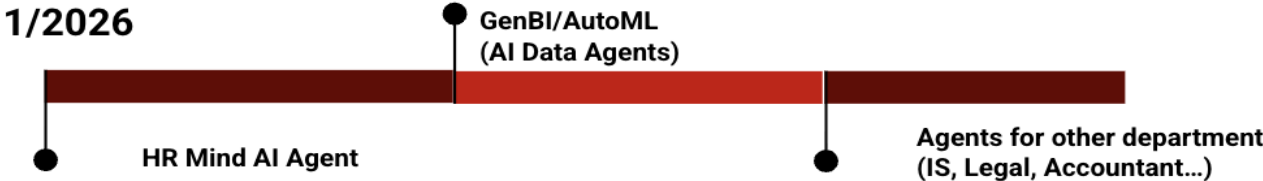


Roadmap 2025-2026

AI Customer Roadmap



AI Internal Roadmap (*)



(*) Thay đổi tùy độ ưu tiên

CHI PHÍ HẠ TẦNG CÔNG NGHỆ (DỰ KIẾN)

Môi trường phát triển:
Chi phí ban đầu: **1,3 – 1.5 tỷ VND**
Chi phí duy trì: **180 triệu VND/năm**

Môi trường chạy thực tế:
Chi phí ban đầu: **2,3 – 2,5 tỷ VND**
Chi phí duy trì: **300 triệu VND/năm**

Giai đoạn đầu thử nghiệm và chờ thiết bị:
Chi phí thuê cloud ban đầu : **1,2 tỷ VND**

Chi phí khác :
800 triệu VND

Chi phí 3 năm đầu
6,8 tỷ VND

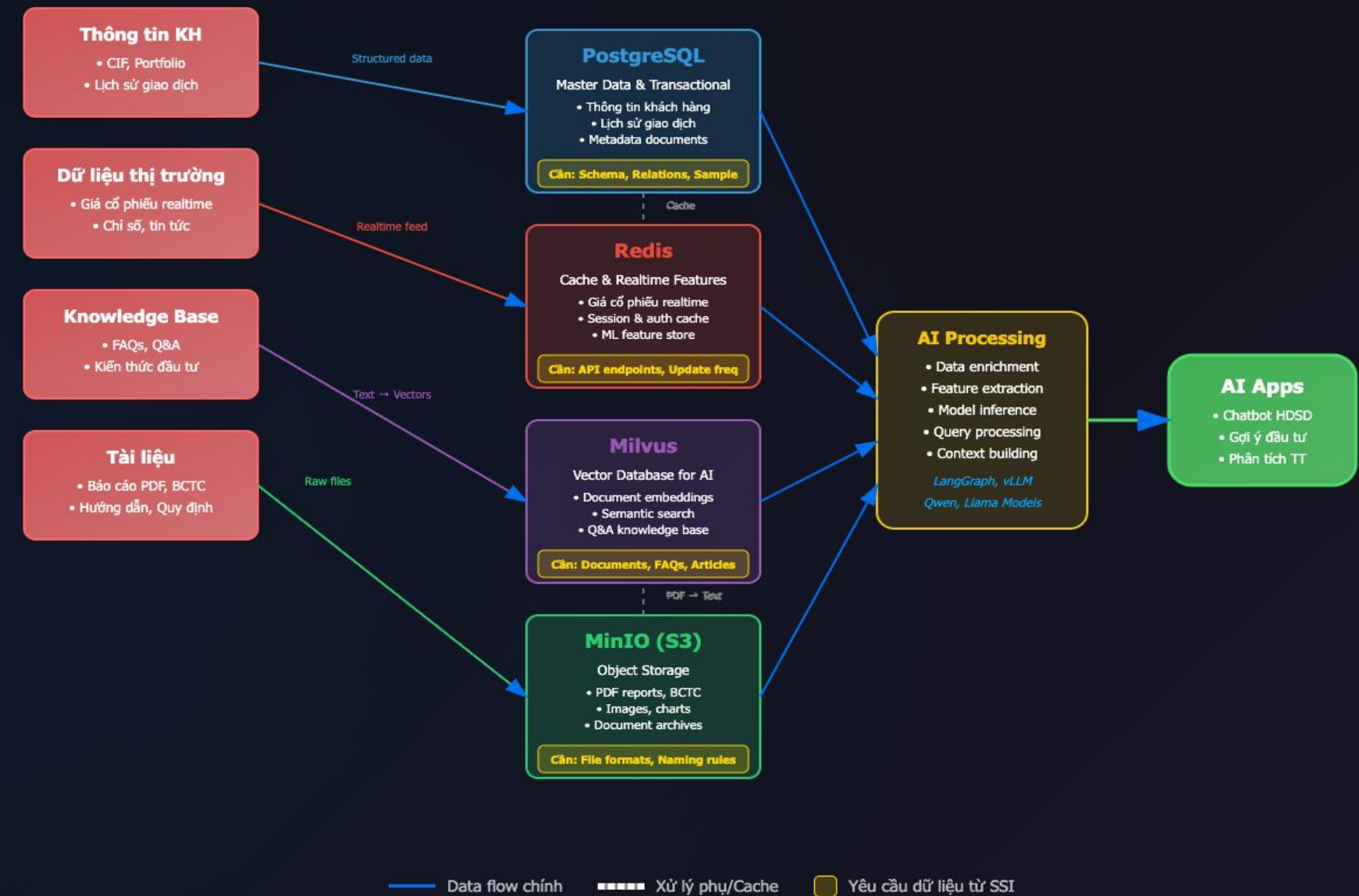
- Note: Cấu hình hiện tại được dự đoán sẽ đủ để vận hành các mô hình mới trong thời gian ít nhất từ 5-6 năm tới, do các mô hình AI có xu hướng nhỏ hơn theo thời gian.



THANK YOU

Database Architecture & Data Flow

Chi tiết nghiệp vụ từng database và yêu cầu dữ liệu từ SSI



Các mô hình AI được tối ưu cho từng tác vụ cụ thể

Document Extraction

SmolDocling-256M

Mô hình nhẹ, tối ưu cho việc trích xuất thông tin từ documents

- › Xử lý PDF, DOCX, HTML
- › OCR và bóc tách tables
- › Trích xuất metadata
- › Tốc độ xử lý nhanh

Use cases: Báo cáo BCTC, Hợp đồng, Tài liệu pháp lý

Large Language Model

Qwen2.5-14B

Mô hình ngôn ngữ mạnh mẽ cho các tác vụ phức tạp

- › Chatbot thông minh
- › Phân tích & tư vấn đầu tư
- › Tạo báo cáo tự động
- › Q&A chuyên sâu

Use cases: Chatbot HDSD, Gợi ý đầu tư, Phân tích thị trường

Vision Language Model

Qwen2.5-7B-VL

Hiểu và phân tích cả văn bản và hình ảnh

- › Phân tích biểu đồ, charts
- › Đọc screenshots
- › Xử lý documents scan
- › Visual Q&A

Use cases: Phân tích biểu đồ kỹ thuật, Báo cáo visual

Embeddings Model

Jina-v4

Vector embeddings cho text và images

- › Semantic search
- › Similar content matching
- › Multi-modal embeddings
- › RAG optimization

Use cases: Tìm kiếm tài liệu, Gợi ý nội dung tương tự

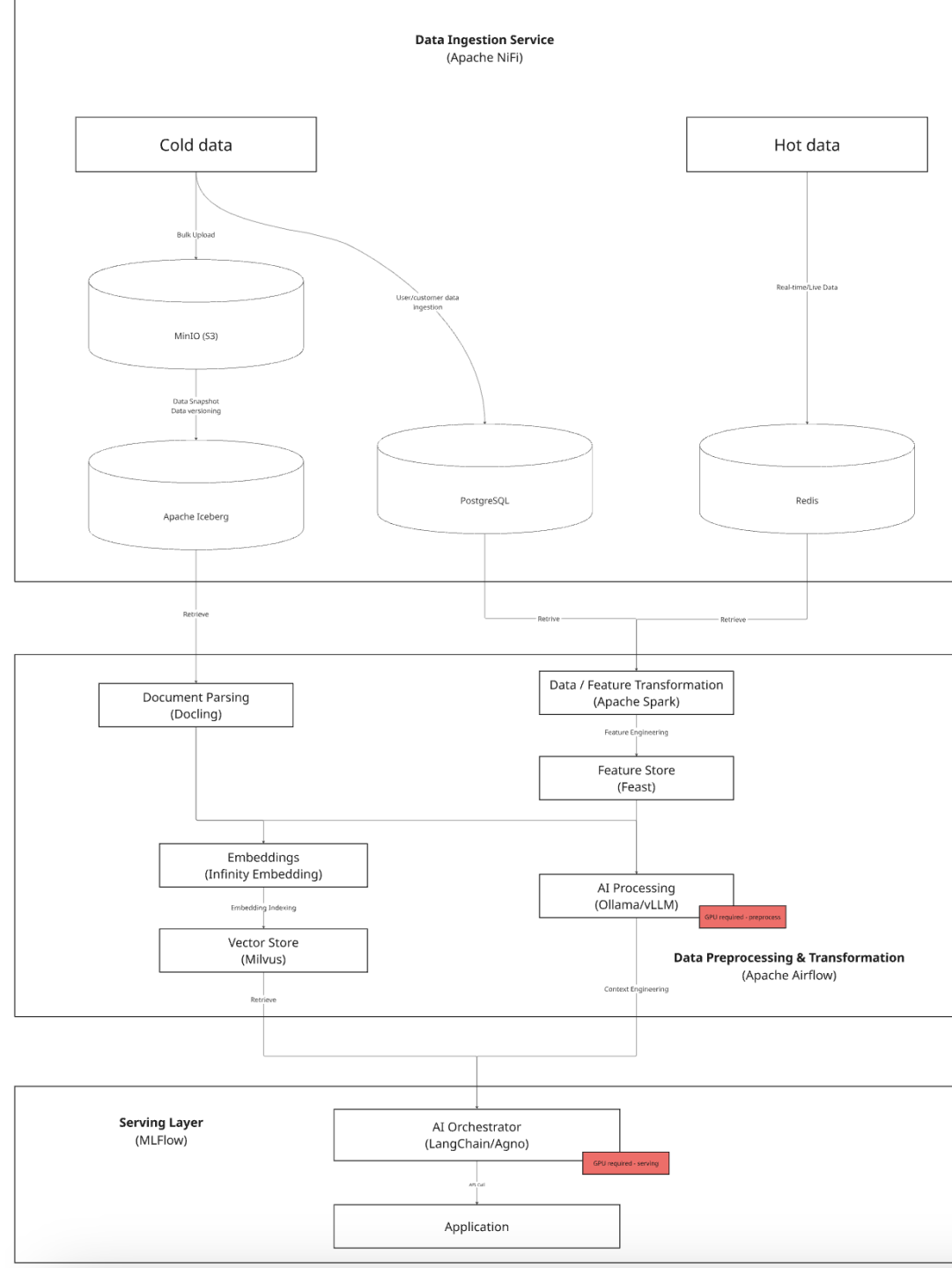
Integration Flow

1. Extract

2. Embed

3. Process

4. Deliver



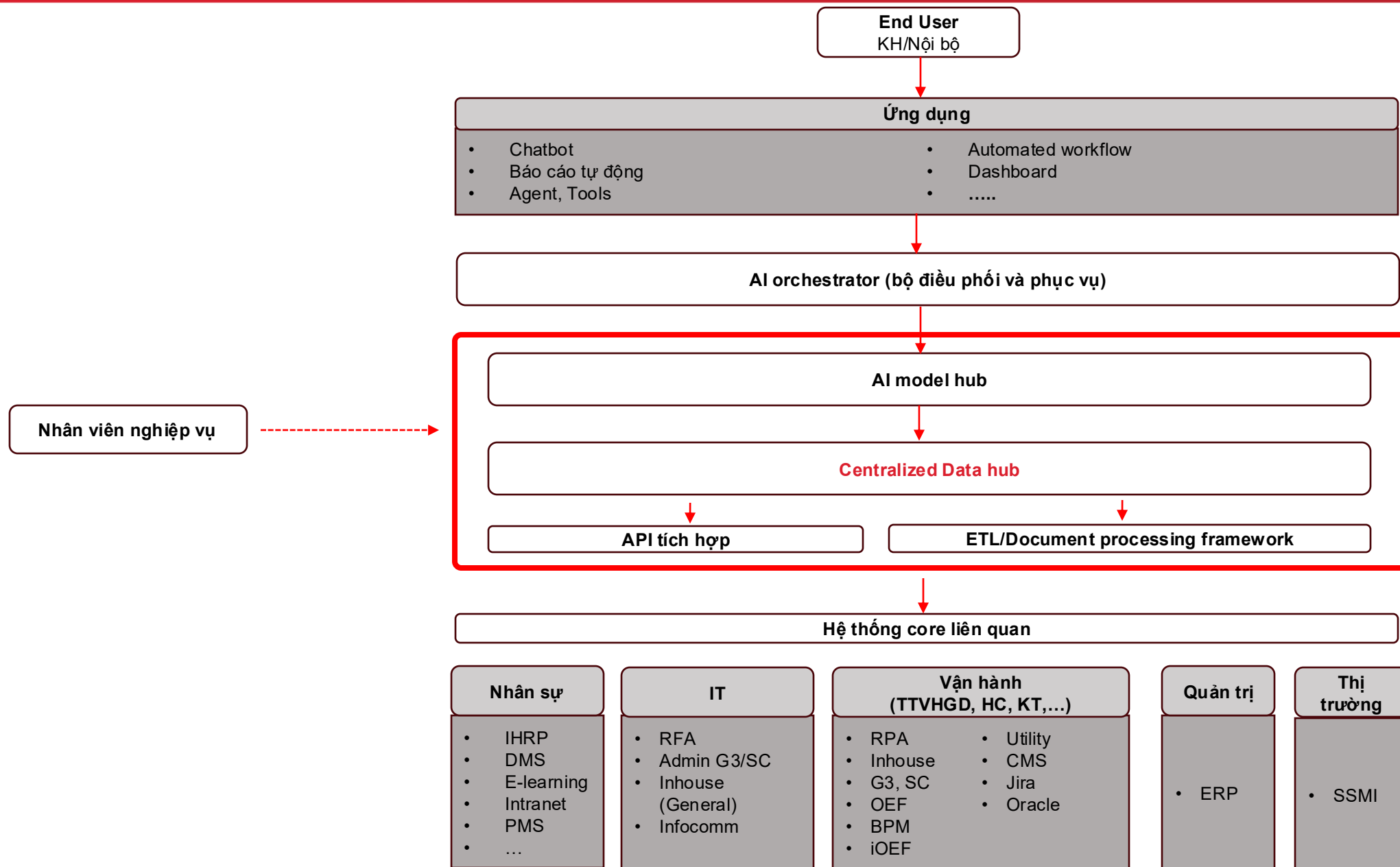
Development Server

Component	Specification
Chassis / Motherboard	ASUS ESC8000A-E12 (4U, dual AMD EPYC 9004, PCIe 5.0)
CPU	Dual AMD EPYC 9004-series
GPU	4 × NVIDIA L40/L40S
RAM	1TB DDR5
Storage	12 TB NVMe (e.g., 6 × 2 TB SSDs)
Power Supply (PSU)	Four 3000 W Titanium redundant PSUs

Production Server

Component	Specification
Chassis / Motherboard	ASUS ESC8000A-E12 (4U, dual AMD EPYC 9004, PCIe 5.0)
CPU	Dual AMD EPYC 9004-series
GPU	2 × NVIDIA H100
RAM	768GB DDR5
Storage	12 TB NVMe (e.g., 6 × 2 TB SSDs)
Power Supply (PSU)	Four 3000 W Titanium redundant PSUs

- Note: Cấu hình hiện tại được dự đoán sẽ đủ để vận hành các mô hình mới trong thời gian ít nhất từ 5-6 năm tới, do các mô hình AI có xu hướng nhỏ hơn theo thời gian.



Chuyển từ Cloud sang On-Premise là quyết định chiến lược mang lại lợi ích lâu dài cho SSI



Bảo mật tuyệt đối

Tuân thủ Nghị Định 13

- Dữ liệu khách hàng không rời khỏi SSI
- Kiểm soát hoàn toàn quy trình bảo mật
- Thoải mái phát triển AI truy cập dữ liệu nhạy cảm
- Đáp ứng yêu cầu tuân thủ ngành tài chính



Tối ưu chi phí

Lợi nhuận dài hạn

- Chi phí cloud tăng theo mức sử dụng
- On-prem: chi phí cố định, sử dụng không giới hạn
- Hoá vốn khi hệ sinh thái AI đủ lớn
- Tiết kiệm đáng kể trong dài hạn



Linh hoạt kỹ thuật

Tùy chỉnh tối đa

- Tích chỉnh nhiều LLM chuyên biệt
- Tối ưu từng mô hình cho tác vụ riêng
- Không bị giới hạn bởi API bên thứ ba
- Kiểm soát hoàn toàn quy trình AI

Từ hiện tại ...

- Chatbot Cú Doha trên cloud
- Hệ thống dashboard cổ phiếu theo TA trên cloud
- Insight MVP 1 + 2 trên cloud
- Dữ liệu phân tán trên nhiều hệ thống

Đến mục tiêu ...

- Toàn bộ hệ thống AI on-premise
- Dữ liệu thống nhất, bảo mật tuyệt đối
- Kiểm soát hoàn toàn mô hình AI
- Tích hợp sâu với hệ thống SSI

Lộ trình chuyển đổi theo giai đoạn

Phase 1: Build stack AI on-prem

Xây dựng nền tảng AI linh hoạt đáp ứng các dự án tại SSI

Phase 2: Chuyển hạ tầng về on-prem

- Chatbot Cú Doha
- TA rating + forecasting

Phase 3: Build iInsight MVP 3 on-prem

Phase 4: Build các chatbot nội bộ

Triển khai chatbot hỗ trợ các bộ phận nội bộ

Xây dựng kho dữ liệu tập trung "AI-Ready" để tăng tốc phát triển ứng dụng AI tại SSI



Tập trung hóa dữ liệu

- Hợp nhất dữ liệu từ nhiều nguồn vào một nền tảng
- Chuẩn hóa format và cấu trúc dữ liệu
- Loại bỏ trùng lặp và đảm bảo tính nhất quán



AI-Ready Data

- Dữ liệu đã được xử lý và làm sạch
- Sẵn sàng cho training và inference
- Tối ưu cho các mô hình AI khác nhau



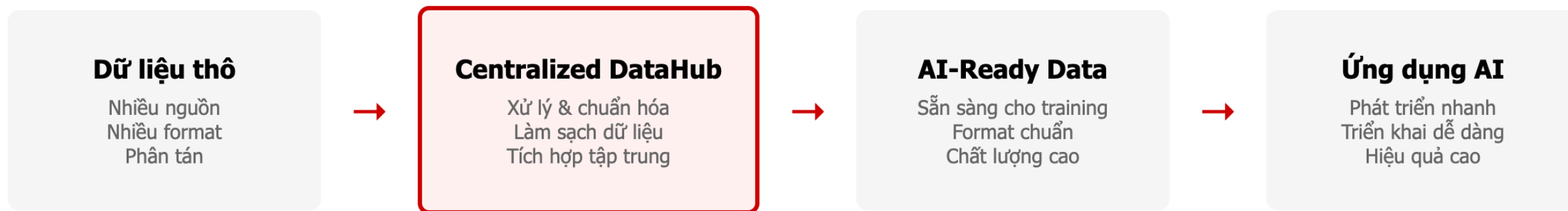
Giảm thời gian phát triển

- Rút ngắn thời gian chuẩn bị dữ liệu
- Tái sử dụng data pipeline đã có
- Tập trung vào phát triển tính năng



Phục vụ đa dạng ứng dụng

- Hỗ trợ cả ứng dụng khách hàng và nội bộ
- Dễ dàng mở rộng cho ứng dụng mới
- API thống nhất cho mọi ứng dụng AI



- Khi phát sinh một ứng dụng AI được phê duyệt triển khai, những dữ liệu cần thiết sẽ được kéo về data hub, và AI model phù hợp sẽ được triển khai thêm
- Dữ liệu trong hub được tích lũy dần qua thời gian, sẽ được sử dụng để phục vụ các ứng dụng AI mới -> Dữ liệu được lưu trữ và chuẩn hóa, tập trung, giảm thiểu việc trùng lặp và giảm thời gian phát triển ứng dụng mới

Data Sources

Thông tin KH

- CIF, Portfolio
- Lịch sử giao dịch

Dữ liệu thị trường

- Giá cổ phiếu realtime
- Chỉ số, tin tức

Knowledge Base

- FAQs, Q&A
- Kiến thức đầu tư

Tài liệu

- Báo cáo PDF, BCTC
- Hướng dẫn, Quy định

Storage Systems

PostgreSQL

- Thông tin khách hàng
- Lịch sử giao dịch
- Metadata

Cần: Schema, Relations

Redis

- Giá realtime
- Session cache
- Quick lookup

Cần: API endpoints

Milvus

- Embeddings
- Semantic search
- Q&A base

Cần: Documents, FAQs

MinIO (S3)

- PDF reports
- Images
- Archives

Cần: File formats



Quy trình triển khai ứng dụng AI on-prem

Usecase: iInsight

Tầm nhìn

- Cố vấn đầu tư AI 24/7
- Khuyến nghị cá nhân hóa
- Dự báo realtime
- Chiến lược điều chỉnh rủi ro

Tính năng chính

- Dự báo giá & xu hướng
- Tối ưu danh mục
- Cảnh báo rủi ro
- Phân tích tâm lý thị trường

Công nghệ

- Deep Learning
- Data pipeline realtime
- Hybrid Cloud + On-prem
- Đa kênh phân phối

1

Giai đoạn 1

Nền tảng dữ liệu

Xây dựng hạ tầng realtime

2

Giai đoạn 2

Mô hình dự báo

Phát triển AI dự đoán

3

Giai đoạn 3

Ra mắt iInsight

Triển khai tư vấn AI

Lưu trữ trong DataHub, phục vụ đa dự án

Lưu trữ trong AI model hub, phục vụ đa dự án

Xây dựng kho dữ liệu AI dự báo

Tích hợp đa nguồn dữ liệu cho AI phân tích và dự đoán chính xác

Dữ liệu chất lượng → Dự báo chính xác

Dữ liệu thị trường

- Giá & khối lượng tick
- Sổ lệnh depth
- Chỉ báo kỹ thuật

Dữ liệu người dùng

- Lịch sử giao dịch
- Hồ sơ rủi ro
- Danh mục đầu tư

Tin tức & tâm lý

- Tin tài chính realtime
- Social sentiment
- Báo cáo phân tích

Dữ liệu cơ bản

- Báo cáo tài chính
- Sự kiện doanh nghiệp
- Chỉ số ngành

Phát triển AI dự đoán chuyên sâu

Xây dựng hệ thống AI dự báo giá, xu hướng và rủi ro cho TTCK Việt Nam

AI học từ quá khứ → Dự báo tương lai

Dự báo giá

- Mô hình LSTM/Transformer
- Dự báo ngắn & trung hạn
- Khoảng tin cậy

Phân tích xu hướng

- AI nhận diện mẫu hình
- Hỗ trợ/Kháng cự
- Chỉ báo động lượng

Đánh giá rủi ro

- Tính toán VaR
- Dự báo biến động
- Phát hiện bất thường

Tối ưu danh mục

- Markowitz optimization
- Lợi nhuận điều chỉnh rủi ro
- Gợi ý tái cân bằng

Development Environment

Chức năng chính:

- Training mô hình AI
- Inference testing
- Document processing

GPU Requirements:

- **Training:** 1 x NVIDIA A100
- **Inference & Processing:** 2 x NVIDIA A6000 (1-2 cards khởi đầu)

Production Environment

Chức năng chính:

- Inference cho ứng dụng thực tế
- Phục vụ khách hàng 24/7
- High availability & performance

GPU Requirements:

- **Inference:** 4 (hoặc 8) x NVIDIA A6000
- **Scalable:** Mở rộng theo nhu cầu

Hạ tầng chung cho cả 2 môi trường

- **PostgreSQL:** Dữ liệu cấu trúc
- **MinIO:** Lưu trữ tài liệu
- **Monitoring:** 24/7 system health
- **Milvus:** Vector search
- **Redis:** Cache & real-time data
- **Backup:** Disaster recovery

✨ Triển khai tư vấn đầu tư AI thể hệ mới

Nền tảng AI tư vấn cá nhân hóa, dự báo chính xác và quản trị rủi ro thông minh

AI Predictive → Cố vấn đầu tư cá nhân 24/7

💡 Khuyến nghị thông minh

- Tín hiệu mua/bán realtime
- Điểm vào/thoát
- Gợi ý cắt lỗ

📊 Phân tích thị trường

- Dự báo thị trường hàng ngày
- Luân chuyển ngành
- Cổ phiếu nóng

⚠️ Quản lý rủi ro

- Cảnh báo danh mục
- Bảo vệ vốn
- Cảnh báo biến động

📱 Đa kênh

- Tích hợp mobile app
- Web dashboard
- API cho đối tác

Thành phần	Cấu hình	Giá (VND)
ASUS ESC8000A-E12 Chassis/Motherboard	4U dual AMD EPYC 9004, PCIe 5.0	~250,000,000
AMD EPYC 9004-series CPU	Dual AMD EPYC 9004-series	~145,000,000
NVIDIA L40/L40s GPU (4 units for dev)	4 x NVIDIA L40/L40s	180,000,000 x 4 = 512,000,000
DDR5 64GB RDIMM	512GB DDR5	10,990,000 x 8 = 87,920,000
PSU 3000W Titanium	4 x 3000W Titanium redundant PSUs	~20,000,000 x 4 = ~80,000,000
NVMe SSD 2TB	TB NVMe (6x2TB SSDs)	40,000,000 x 6 = 240,000,000
Các thiết bị khác (NIC , module quang, dây cáp quang, dây cáp mạng, HBA)		50,000,000 – 60,000,000

Môi trường phát triển:
Chi phí ban đầu: **1,3 – 1.5 tỷ VND**
Chi phí duy trì: **180 triệu VND/năm**

- Note: Cấu hình hiện tại được dự đoán sẽ đủ để vận hành các mô hình mới trong thời gian ít nhất từ 5-6 năm tới, do các mô hình AI có xu hướng nhỏ hơn theo thời gian.

CHI PHÍ HẠ TẦNG CÔNG NGHỆ (DỰ KIẾN)



Thành phần	Cấu hình	Giá (VND)
ASUS ESC8000A-E12 Chassis/Motherboard	4U dual AMD EPYC 9004, PCIe 5.0	~250,000,000
AMD EPYC 9004-series CPU	Dual AMD EPYC 9004-series	~145,000,000
NVIDIA H100 GPU (2 units for prod)	2 x NVIDIA H100 80GB	747,000,000 x 2 = 1,494,000,000
DDR5 64GB RDIMM	1TB DDR5	10,990,000 x 15 = 164,850,000
PSU 3000W Titanium	4 x 3000W Titanium redundant PSUs	~20,000,000 x 4 = ~80,000,000
NVMe SSD 2TB	TB NVMe (6x2TB SSDs)	40,000,000 x 6 = 240,000,000
Các thiết bị khác (NIC , module quang, dây cáp quang, dây cáp mạng, HBA)		50,000,000 – 60,000,000

Môi trường chạy thực tế:
Chi phí ban đầu: **2,3 – 2,5 tỷ VND**
Chi phí duy trì: **300 triệu VND/năm**

- Note: Cấu hình hiện tại được dự đoán sẽ đủ để vận hành các mô hình mới trong thời gian ít nhất từ 5-6 năm tới, do các mô hình AI có xu hướng nhỏ hơn theo thời gian.

Chi phí hạ tầng AI - Training System

Cấu hình hệ thống Training

STT	Phần cứng	Chi tiết
1	MAIN	ASUS Pro WS WRX80E-SAGE SE WIFI
2	CPU	AMD Ryzen Threadripper Pro 5975WX (32C/64T)
3	HSF	Noctua NH-U14S TR4-SP3
4	RAM	2x ECC REG SK hynix 64GB DDR4 (Tổng: 128GB)
5	HDD	WD Red Plus 4TB (SATA 256MB)
6	SSD	WD Black SN850X 2TB (M2 NVMe 4.0)
7	PSU	Super Flower LEADEx PLATINUM SE 1200W
8	CASE	Segotep Phoenix T1 (BLACK, eATX) + 3 fan

GPU cho Training:

NVIDIA H100 80GB × 2 cards

Chi phí dự tính - Training

Hệ thống máy chủ:	83,900,000 VND
NVIDIA H100 80GB:	720,000,000 VND/card
Tổng GPU (2 cards):	1,440,000,000 VND

TỔNG CHI PHÍ TRAINING

1,523,900,000 VND

(Chưa bao gồm VAT)

Cần tối thiểu 1 GPU từ Giai đoạn 2 để làm huấn luyện các mô hình.
Lưu ý: Giá hiển thị là giá từ nhà cung cấp thứ 3, do hãng không bán trực tiếp ở Việt Nam