THESIS REPORT #3

МИКРОСЕРВИС АРХИТЕКТУРТ СУУРИЛСАН ХИЙМЭЛ ОЮУН АГЕНТУУД

МУИС, МТЭС, МКУТ, Мэдээллийн технологи хөтөлбөр, 4-р түвшний оюутан Б.Раднаабазар

2025 оны 11-р сарын 18

Агуулга

01 Нэр томьёоны тайлбар

Мэдээлэл	Хүний мэдрэхүйн дамжиж хүний ой эрхтэн, ухаан, баримт, эд зүйл,бичиг зайд ямар нэг хадгалж байгаа дижитал орон зайд ямар нэгэн байдлаар үлдэнэ
Мэдээлэл	Тодорхой зорилгод чиглэсэж боловсруулсан, утга санаа илэрхийлсэн өгөгдөл
Өгөгдөл	Ямар нэг баримт (факт), статистик эсвэл мэдээллийн элемент бөгөөд тоон шинжийг агуулна
Эдийн засаг	Бараа, бүтээгдэхүүн (үйлчилгээ) үйлдвэрлэх, худалдах, худалдан авах хэлбэрээр орлого олж, ашиг олох үйл ажиллагаа.
Цахим бизнес	Интернэт, ICT-ийн тусламжтай онлайн орчинд явуулж буй бизнесийн үйл ажиллагаа;
Хиймэл оюун	Компьютерийн системийг хүний сэтгэн бодох, шийдвэр гаргах, суралцах чадвартай болгох технологийн салбар.
Суурь модел (Foundation Model)	Олон төрлийн өгөгдөл дээр урьдчилан сургасан, олон даалгаварт ашиглагдах чадвартай хиймэл оюуны том загвар.
Том хэлний модел (Large Language Model, LLM)	Хүний хэлний бүтэц, утгыг ойлгож, бичвэр үүсгэх чадвартай хиймэл оюуны загвар. Transformer архитектур дээр суурилдаг.

01 Нэр томьёоны тайлбар

Хиймэл оюуны инженерчлэл (AI Engineering)	Бэлтгэгдсэн суурь моделийг ашиглан бодит хэрэглээний аппликейшн, систем хөгжүүлэх хөгжүүлэлтийн арга барил.
Prompt инженерчлэл (Prompt Engineering)	Хиймэл оюуны моделд өгч буй зааварчилгаа, асуултыг оновчтой бичих замаар хүссэн үр дүн гаргуулах арга.
Хиймэл оюун агент (AI Agent)	Өөрийн орчныг мэдрэх, нөхцөл байдлыг ойлгож, төлөвлөгөө гарган, хэрэгсэл ашиглан үйлдэл хийдэг ухаалаг систем.
Агентын төлөвлөлт (Planning)	Агент даалгавраа биелүүлэхийн тулд дараалсан алхамуудыг тодорхойлох үйл явц. Жишээ: зорилго → төлөвлөгөө → гүйцэтгэл → дүн шинжилгээ.
RAG — Хайлтаар нэмэгдүүлсэн үүсгэлт	Хиймэл оюуны модел гадаад өгөгдлийн сангаас холбогдох өгөгдлийг хайж, хариулт гаргахдаа ашигладаг арга.
Embedding (Вектор дүрслэл)	Текст эсвэл өгөгдлийг тоон вектор хэлбэрт хувиргах арга. Семантик утгаар нь ойролцоо өгөгдлийг хайхад ашигладаг.
Микросервис архитектур	Том системийг жижиг, бие даасан, тусдаа ажиллах сервисүүдэд хуваах хөгжүүлдэг программ хангамжийн архитектур.
Монолит архитектур	Бүх функц, логик, өгөгдлийн сан нь нэг програмд нэгтгэгдсэн уламжлалт программ хангамжийн архитектур.

01 Оршил

- AI нь компаниудын заавал нэвтрүүлэх ёстой бай болж байна. (Harvard Review, 2023)
- 2024 онд S&P 500 Компаниуд AI нэвтрүүлэх нь өмнөх жилийхээс 3 дахин их ихэслээ.
- AI engineering нь хамгийн хурдацтай өсөх бизнес боллоо. (Github дээр stars-ийн тоогоор тэргүүлж байна.)

Гол асуудал:

Хиймэл оюуны агент нь өргөнөөр хэрэгжиж байна → **Монолит агентууд нь NxM нягт холбоостой байдаг** → Нэг агент унах нь дамжин унах алдаа үүсгэдэг → **Өргөжүүлэх, засварлахад төвөгтэй. Бизнесийн үйл ажиллагаа удаашралтай**

Huyen, C. (2024). AI Engineering. O’Reilly Media.

01 Судалгааны зорилго

Хиймэл оюун агентуудыг микросервис архитектурт нэвтрүүлж, уян хатан, өргөжүүлэх боломжтой, найдвартай систем бүтээх

Судалгааны зорилт

1. Хиймэл оюуны инженерчлэлийн онол, суурь модел, RAG системийг судлах
2. Хиймэл оюун агентуудын архитектур, төлөвлөлтийн механизм судлах
3. Микросервис архитектурын давуу, сул талуудыг тодорхойлох
4. Агентуудыг монолит архитектурт нэвтрүүлэхэд тулгарах асуудлуудыг тодорхойлох
5. Kafka-Flink ашиглан EDA суурилсан архитектур санал болгох
6. Практик демо систем хөгжүүлж туршиж үзэх

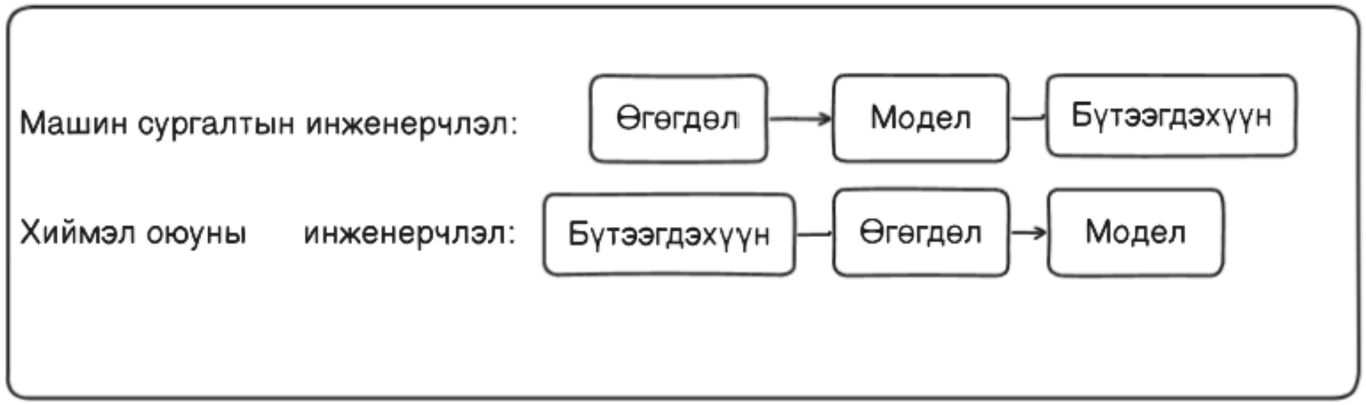
01 Хиймэл оюуны инженерчлэл

Уламжлалт ML

- Модел хөгжүүлэх
- Өгөгдөл цуглуулах
- Математик мэдлэг шаардлагатай

Хиймэл оюуны инженерчлэл

- Бэлтгэгдсэн модел ашиглах (суурь, жижиг хэлний модел...)
- Программ хангамжид интеграци хийх
- Зааврын инженер хийх



Huyen, C. (2024). AI Engineering. O'Reilly Media.

01 Хиймэл оюуны хөгжил

Хөгжлийн замнал

- Хэл модел: Статик өгөгдөл рүү кодлох
- Том хэлний модел (LLM): Тэрбум гаруй параметртэй
- Суурь модел: Мульти модал, ерөнхий зориулалттатай

- Transformer архитектур (2017)
- Өөрийгөө удирдсан сургалт
- GPT-3: 175 тэрбум параметр
- GPT-4: 1.2 их наяд параметр

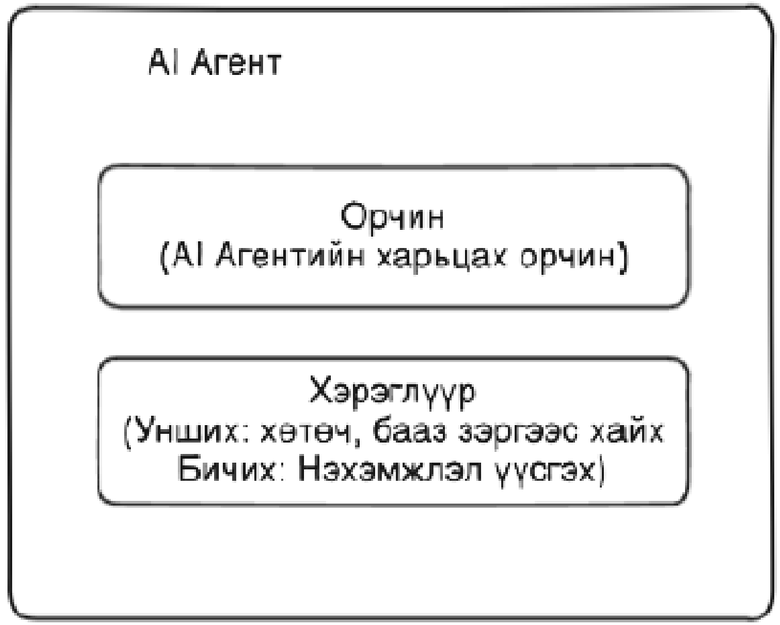
- Attention механизм
- Урьдчилсан сургалт
- Дараах сургалт
- Sampling стратеги

Huyen, C. (2024). AI Engineering. O’Reilly Media.

01 Хиймэл оюуны агент

Агент гэж юу вэ?

Өөрийн орчныг мэдрэх, түүн дээр үйлдэл хийх чадвартай систем



Huyen, C. (2024). AI Engineering. O’Reilly Media.

Үргэлжлэл

Хиймэл оюунд гадна орчинтой хандах хэрэглүүр өгөх нь маш олон боломжийг нээж өгдөг. Даалгаврын дагуу үйл ажиллагааг зохион байгуулж хэрэгжүүлдэг программ хангамж юм.

- Орчин: Ажиллах орчин (Гал тогоо, интернет, дотоод систем)
- Үйлдэл: Хийж чадах үйлдлүүд (цаг захиалах, тооцоолол хийх)
- Даалгавар: Хэрэглэгчийн хүсэлт (хэрэглэгчийн өмнөөс нэхэмжлэл гаргах, шинжилгээ хийх)

01 Хайлтаар нэмэгдүүлсэн үүсгэлт (RAG)

RAG гэж юу вэ?

Моделийн мэдлэгийг гадаад эх сурвалжаар өргөтгөж, төөрөгдлийг багасгаж, найдвартай хариулт өгдөг систем

Хоёр гол бүрэлдэхүүн

- Хайгч: Холбогдолтой баримт олох
- Үүсгэгч: Баримт ашиглан хариулт үүсгэх

RAG-ийг байнга үнэлэх шаардлагатай

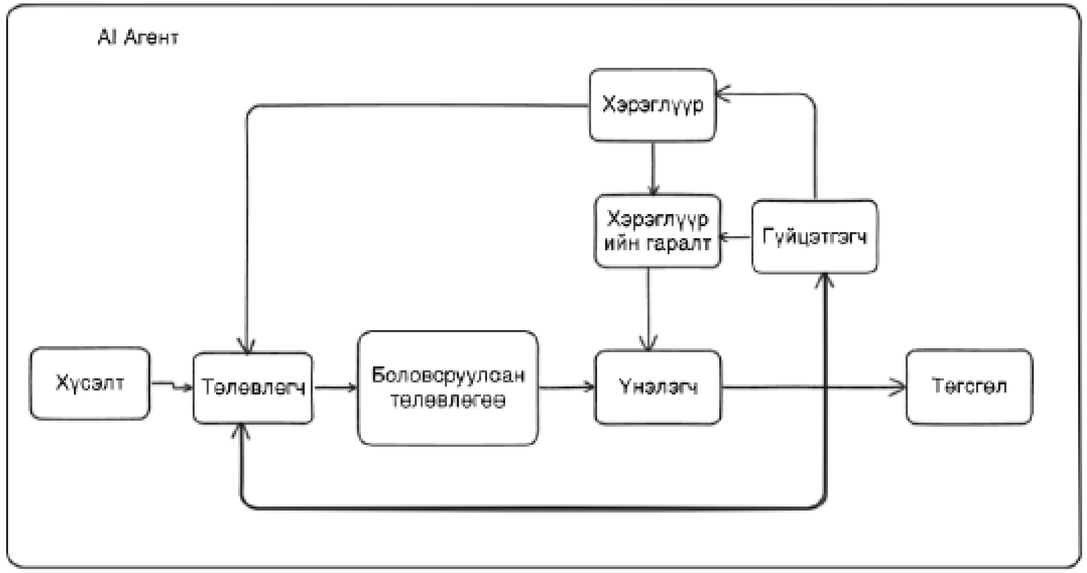
Өгөгдөл тогтмол ихэсдэг учир тасралтгүй хөгжүүлэлт хийснээр бодит хэрэглээнд үр нөлөө алдахгүй байх боломжийг бүрдүүлнэ

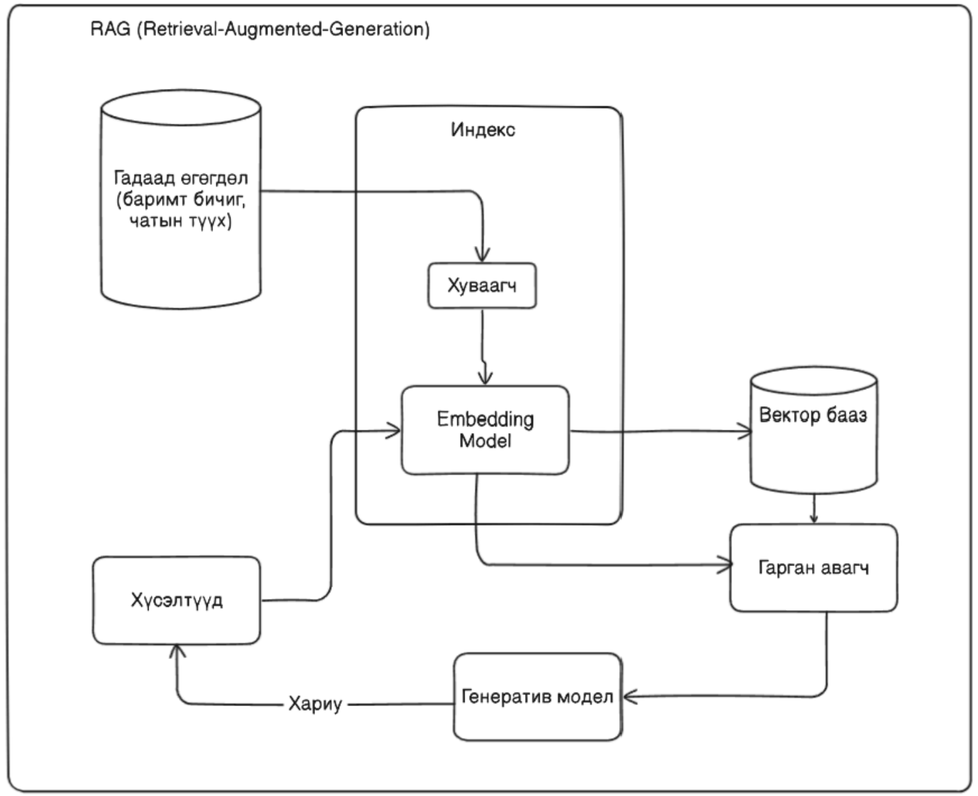
Хайлтын аргууд

- TF-IDF (нэр томьёо)
- Вектор хайлт (утга зүй)
- FAISS, Milvus, Pinecone (векторлохдоо граф, мод бүтэ ашиглах)

Huyen, C. (2024). AI Engineering. O’Reilly Media.

01 RAG болон хиймэл оюун агент зохиомж





02 Микросервис

Монолит

- Нэг том систем
- Нэг кодын сан
- Хялбар хөгжүүлэлт
- Өргөжүүлэх хүндрэлтэй

Микросервис

- Нэг том систем
- Нэг кодын сан
- Хялбар хөгжүүлэлт
- Өргөжүүлэх хүндрэлтэй

Сорилт ба хиймэл оюун агентууд

Монолит нь бүгдийг нэтгэсэн байдлаар өргөжүүлдэг. Программ хангамжийн бүрэлдэхүүнүүд нягт хамаарал бүхий харилцаагаар холбогдож, алдаа нэмдэг. Иймээс хиймэл оюуны агентийн процессийн урсгалыг монолитээр хэрэгжүүлэх нь монолит системийн тулгардаг асуудлууд гарч байдаг.

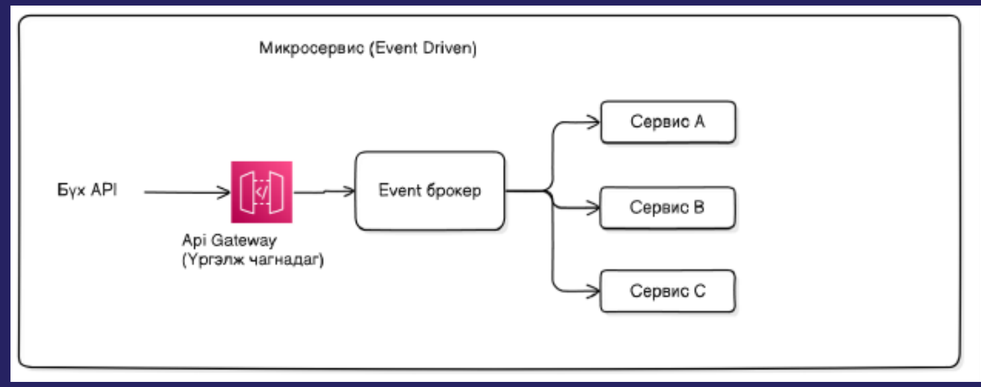
Fowler, Martin (2014). Microservices: A Definition of This New ArchitecturalTerm

02 Үйл явдлаар удирдагдах архитектур (EDA)

Синхрон харилцаа

- HTTP REST / gRPC
- NxM нягт холбоос
- Нэг унавал бүгд дамжин унана

Асинхрон харилцаа



- Мессежийн брокер
- Салангид байдал
- Найдвартай уян хатан байдал

Falconer, Sean (2025). The Future of AI Agents is Event-Driven

02 EDA ба хиймэл оюун агентууд

Apache Kafka

Apache Kafka нь салангид, өндөр дамжуулалттай, бага хоцрогдолтой EDA-ийн төв мэдээллийн систем болж чаддаг. Лог хадгалах зарчмаар ажилладаг учир үнэлэхэд хүчтэй технологи болно.

Kafka-ийн үндсэн бүрэлдэхүүнүүд

- Topic: Үйл явдлуудыг ангилах суваг
- Producer:Үйл явдал үүсгэгч
- Consumer: Үйл явдал хүлээн авагч
- Partition: Өргөжүүлэх, параллель боловсруулалт
- Consumer: Group Ачааллыг хуваарилах

Flink-ийн давуу талууд

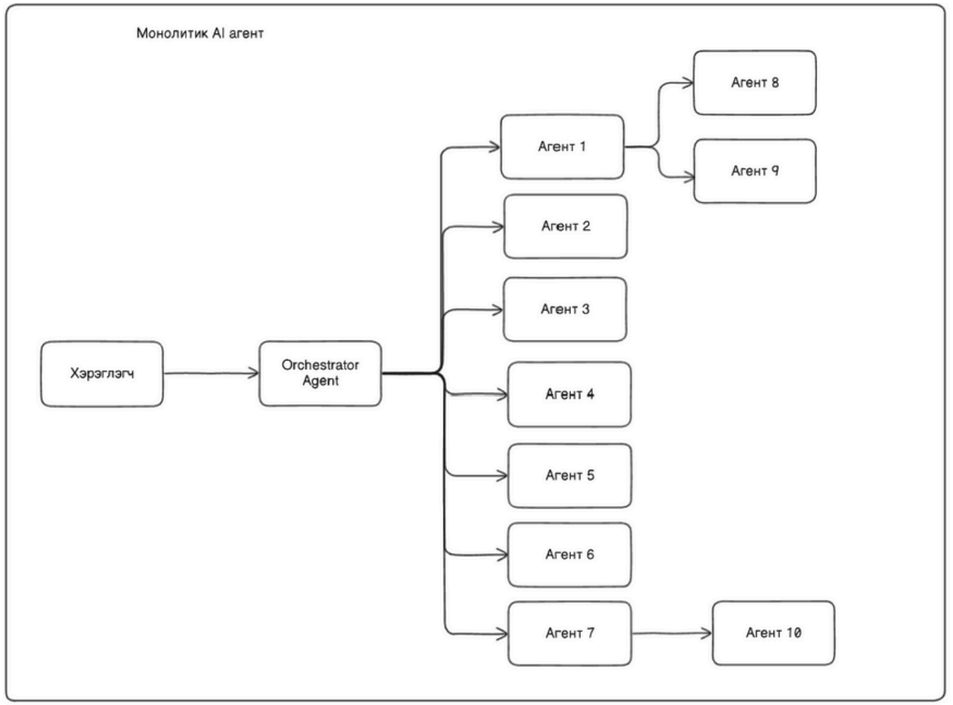
- Төлөв байдлын удирдлага - Найдвартай, нарийн тооцоолол
- Бодит цагийн боловсруулалт - Үйл явдлын цаг дээр суурилсан
- Өндөр дамжуулалт - Секундэд сая сая үйл явдал
- ”Яг нэг” семантик - Мэдээлэл яг нэг удаа боловсруулагдана

Flink ба хиймэл оюун

Flink нь LLM-тэй холбогдож, төлөвлөгч агентыг хэвтээ өргөжих боломжийг олгодог. RAG системийг Flink-тай хослуулж, бодит цагт өгөгдөл боловсруулж, моделийн мэдлэгийг тасралтгүй шинэчилдэг.

Falconer, Sean (2025). The Future of AI Agents is Event-Driven

03 Асуудал ба шийдлийн зохиомж



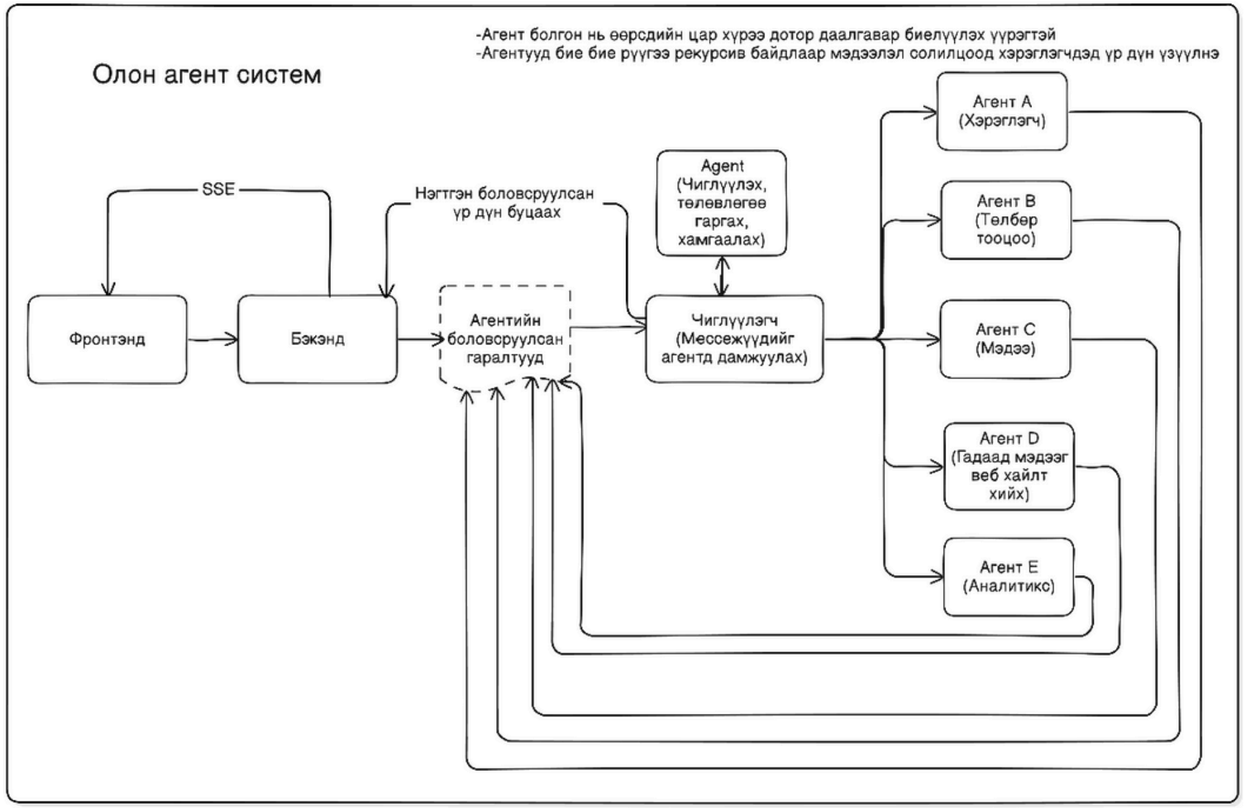
Монолит агентуудын эрсдэл

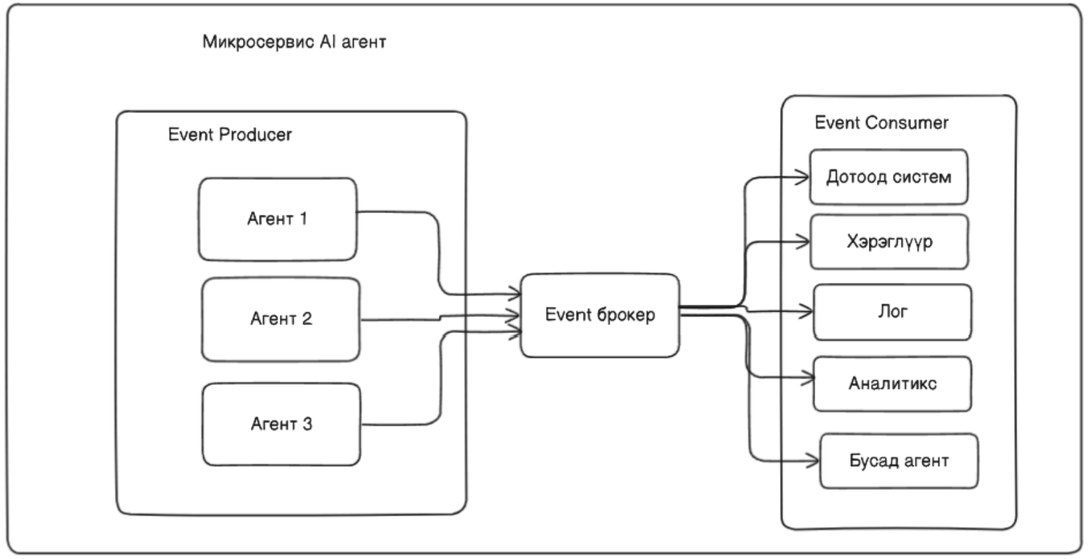
- NxM нягт холбоос - Өргөжүүлэх хүндрэлтэй
- Дамжин унах алдаа - Нэг агент унавал бүх систем унана
- Хөгжүүлэлтийн удаашрал - Бүх системийг дахин суурилуулах
- Тасралтгүй арчилгаа- дэд процессийн үйл ажиллагаа өөрчлөгдвөл алдаа гарах, нэг газар унавал алдаа барихад хэцүү байх

Олон агентийн шаардлага

Бизнесийн цар хүрээнүүдийн үйл ажиллагаанууд ялгаатай байдаг тул олон агент хэрэг болдог. Жишээ нь үнэтэй тооцоолуурын агент, хэрэглэгчийн зөвлөх агент, чиглүүлэгч агент г.м.

03 Асуудал ба шийдлийн зохиомж





Микросервис архитектурд суурилсан агентуудын зохиомж

- Эвент үүсгэгчүүд - Өөр өөр зориулттай агентууд
- Эвент брокер - Apache Kafka буюу мессеж брокер
- Эвент хүлээн авагч - Агентууд, хэрэглүүрүүд

Ажиллах зарчим

Хэрэглэгч → API Gateway → Kafka → Orchestrator → Агентууд (Мэдээллээ нийлүүлэхийн тулд бусад агентуудыг рекурсив дуудах)→ SSE → Frontend

03 Асуудал ба шийдлийн зохиомж

Техникийн давуу талууд

- Салангид байдал - Агентууд бие биенээсээ хараат бус
- Асинхрон харилцаа - Параллель боловсруулалт (Бодит цагийн боловсруулалт)
- Бие даан өргөжүүлэх - Агент бүр өөрийн хэрэгцээний дагуу
- Найдвартай байдал - Үйл явдлууд хадгалагдана
- Рекурсив ажиллагаа - Агентууд дахин зохион байгуулагч руу хандаж болно

Бизнесийн давуу талууд

Бодит цагийн шийдвэр гаргалт, Лог хадгалалт (аудит, тест), Хэвтээгээр өргөжүүлэх, Шинэ агент нэмэхэд хялбар.

04 Хэрэгжүүлэлт

Демо төсөл

Монгол болон дэлхийн хөрөнгийн зах зээлийн ухаалаг дүн шинжилгээний программ хангамж

Өгөгдлийн эх сурвалж

- МХБ-ийн бодит өгөгдөл
- Finnhub API
- TradingView Widgets

Гол функцууд

- Хэрэглэгчийн бүртгэл (найруулсан эмэйл илгээх), профайл (JWT токен)
- Watchlist үүсгэх, хувьцаа нэмэх/хасах
- Өдөр тутам өөрт тохируулан найруулсан эмэйл авах
- AI агентаар хувьцааны шинжилгээ асуух
- Хувийн портфолионы дүн шинжилгээ авах

04 Хэрэгжүүлэлт

Фронтенд шаардлага

- Figma дизайны дагуу хийх
- Алдааны мессеж харуулах
- Харанху горим
- Responsive байх
- NextJS, TypeScript

Бэкенд шаардлага

- Postgres Native удирдлагын сан
- JWT нэвтрүүлэх
- NodeJS, TypeScript, Zod Type-safety, ExpressJS
- CI/CD, Docker Container ашиглах

Интеграцийн шаардлага

- Кафка, Флинк зэрэг нэвтрүүлэх
- Гадаад дотоод мэдээлэл бодит цагийн горимоор харуулах
- Хиймэл оюун агентуудыг нэвтрүүлэх.
- Демо системийг бодит сервер дээр байршуулах

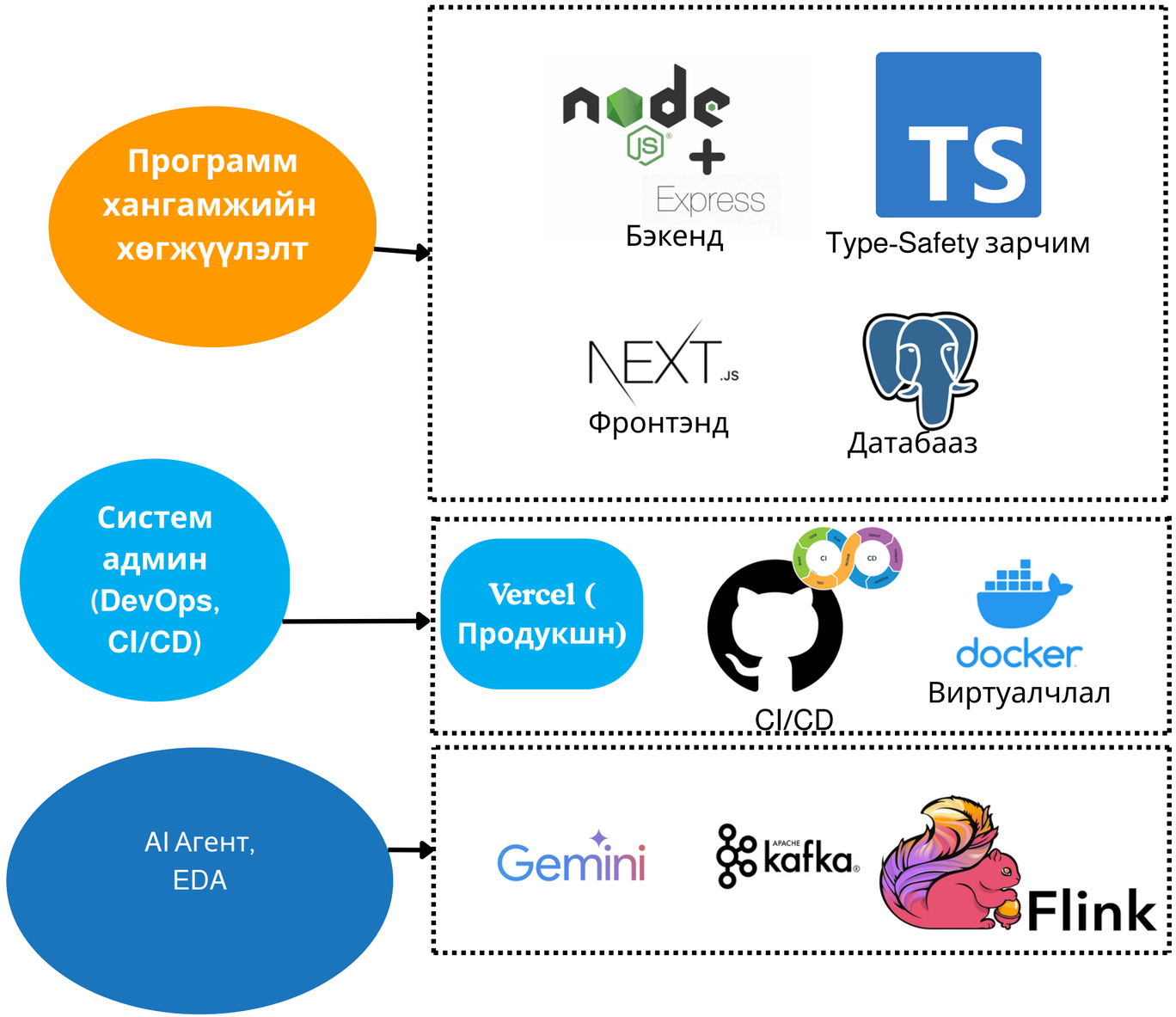
Функциональ шаардлага

- Хэрэглэгчийн тохирсон дүн шинжилгээг эмэйлээр явуулах
- Гадаад, дотоод зах зээлийн мэдээлэл- Бодит цагаар харах, шинжилгээ харах, портфолио удирдах
- Хувийн мэдээлэл засварлах (watchlist)

Функциональ бус шаардлага

- Аюулгүй байдал- HTTPS, JWT, OAUTH
- Хариу үйлдэл- 3000ms дотор харуулах, шаардлагатай бол redis-д кеш хийх
- Хүртээмж- 99 хувийн uptime. reAct.

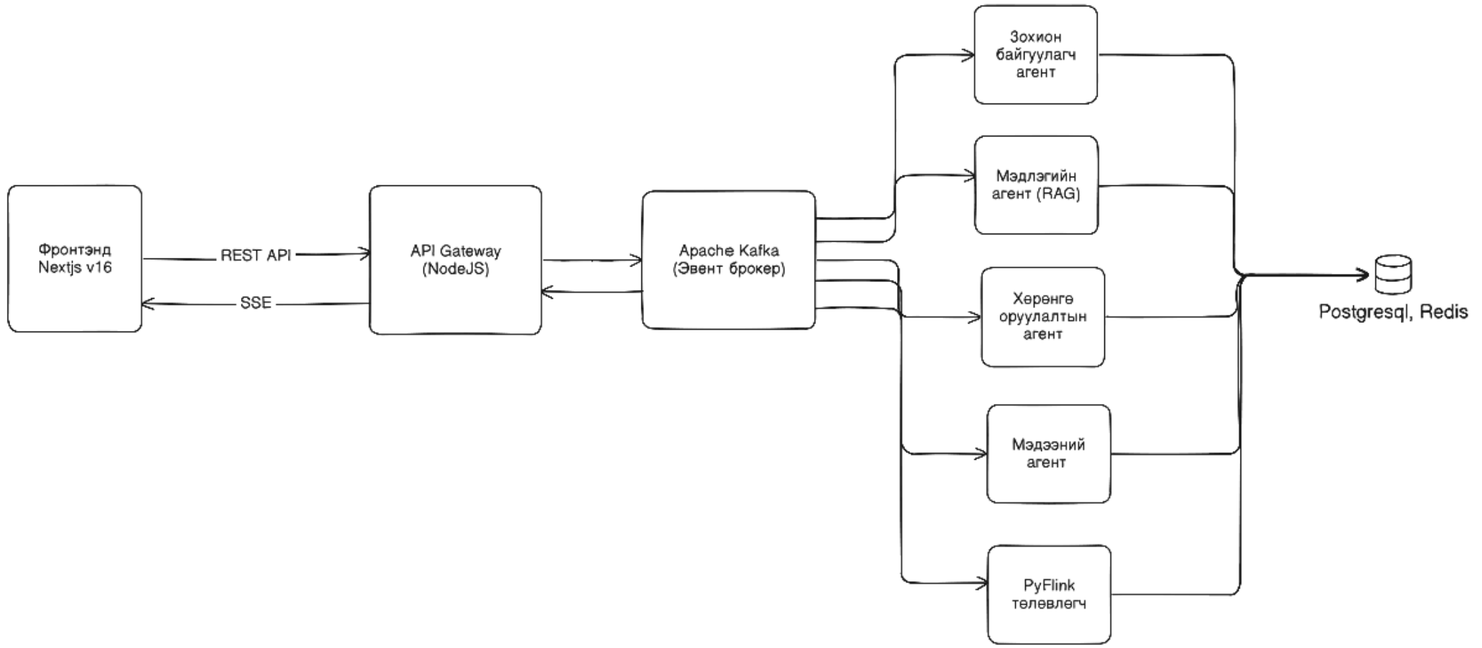
04 Шийдлийн технологийн стек



04 Хэрэгжүүлэлт

Агентууд	Үүрэг
Orchestrator （Зохион байгуулагч агент）	Gemini AI ашиглан хүсэлтүүдийг ангиллаж, агентууд руу даалгавар шилжүүлэх
Knowledge （Мэдлэгийн агент）	RAG систем, vector search ашиглан мэдлэг хайх. Бүх агентуудын метадата, хэрэглүүр зэрэх орох
Investment （Хөрөнгө оруулалтын агент）	МХБ өгөгдөл дүн шинжилж, зөвлөмж өгөх
News （Мэдээний агент）	Finnhub API-ээр мэдээ цуглуулж хураангуйлах
PyFlink Planner	Рекурсив эсвэл жирийн чиглүүлэг хийх

Архитектурын зохиомж



04 Хэрэгжүүлэлт

Хөгжүүлэлтийн хугацаа

- Хугацаа: 6 долоо хоног (Жинхэнэ хамгаалалт хүртэл 3 долоо хоног)
- Хэрэгжүүлэлт: 70% бүрэн

Бүрэн хэрэгжсэн хэсгүүд

- Docker, Kafka, PostgreSQL, Redis суурилуулалт
- Бүх агентууд (Orchestrator, Knowledge, Investment, News, Flink Planner)
- API Gateway
- Frontend (Vercel дээр байршуулсан)

Гүйцэтгэл

- Датабааз хайлт: 50-100мс
- Кафка дамжуулалт: 5-10мс
- API Gateway: 200-500мс
- Хайлтаар нэмэгдүүлсэн үүсгэлт: ~10секунд

04 Хэрэгжүүлэлт

Дэд бүтэц （100%)

- Бүх сервисд Docker Compose суурилуулалт
- Apache Kafka 3.5
- PostgreSQL 16 - pgvector extension
- Redis 7 - кеш
- Kafka сэдвүүд

API Gateway （100%)

- Authentication
- Watchlist CRUD
- AI агенттай харьцах （query, SSE）
- Бүх агентийн health check

Orchestrator Agent (100%)

- Intent classification (6 categories)
- Gemini AI integration
- Dynamic routing
- Request caching
- Monitoring events

Knowledge Agent (70%)

- RAG системд метадата оруулах(all-MiniLM-L6-v2)
- PostgreSQL pgvector

PyFlink Planner (70%)

- Kafka-гийн мэдээллийг нэтгэх, чиглүүлэх
- Төлөв хадгалагддаг байдлаар хэрэгжүүлэх

04 Хэрэгжүүлэлт

Хязгаарлалтууд

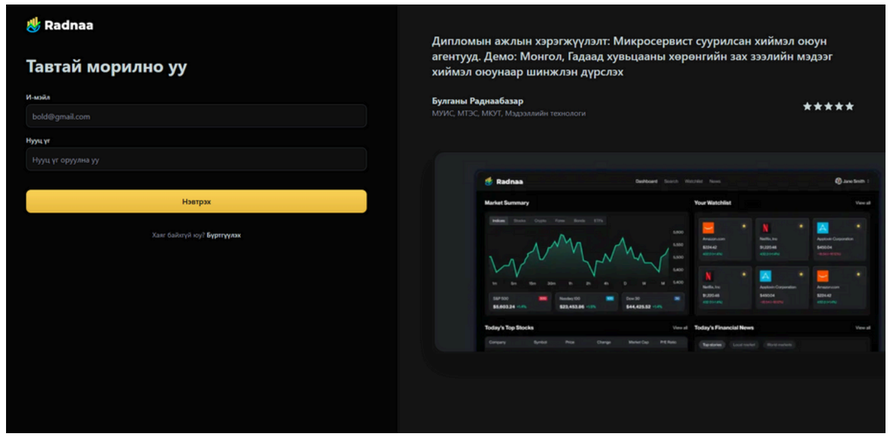
- Демо систем нь үндсэн функцуудыг агуулсан
- Портфолио удирдлага, эрсдлийн үнэлгээ зэрэг нарийн төвөгтэй функцууд хараахан хөгжүүлэгдээгүй

Цаашдын судалгаа

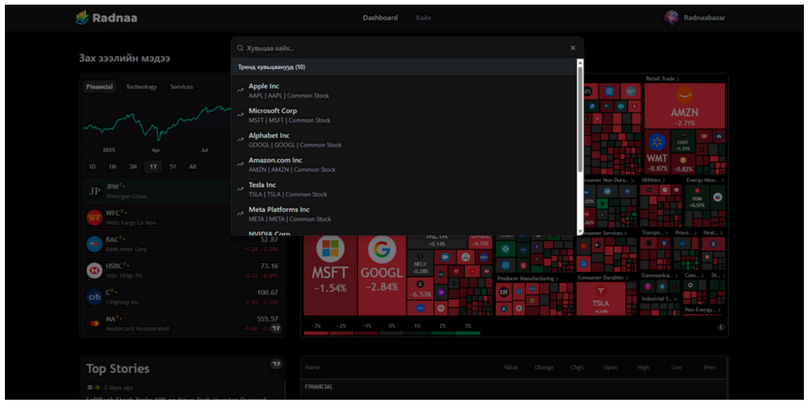
- Production орчинд өргөжүүлэх
- Өндөр ачааллын тест хийх
- Бусад domain дээр хэрэглэх боломжийг судлах

https://stock-tracker-app.vercel.app/

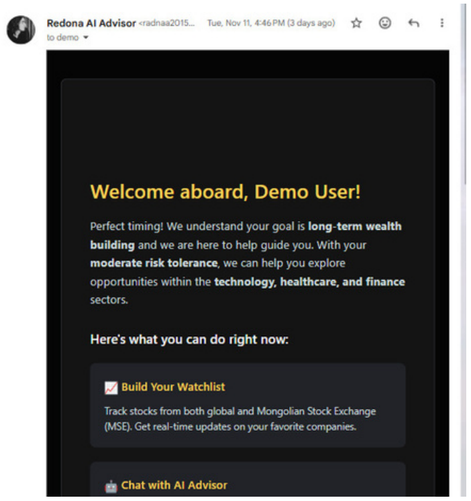
04 Хэрэгжүүлэлт



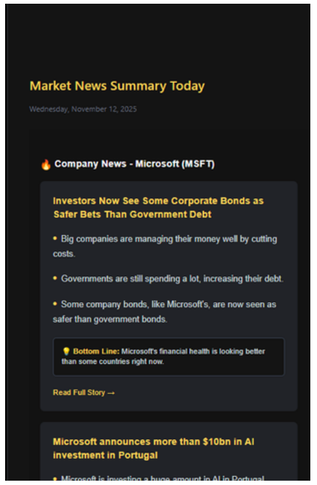
Нэвтрэх хуудас



Нүүр хуудас



бүртгэлийн мэдэгдэл- эмэйл



бүртгэлийн мэдэгдэл- эмэйл

Ном зүй

- Huyen, Chip. AI Engineering. O'Reilly Media, 2024.
- Vaswani, A., et al. "Attention Is All You Need". Advances in Neural Information Processing
- Falconer, Sean. "AI Agents are Microservices with Brains". March 2025.
- Falconer, Sean. "The Future of AI Agents is Event-Driven". BigDataWire, March 2025.

АНХААРАЛ ХАНДУУЛСАНД БАЯРЛАЛАА