

Б.Раднаабазар

Монгол Улсын Их Сургууль
Мэдээллийн Технологи, Электроникийн Сургууль

2025 оны 11 сар



- 1 Удиртгал
- 2 Онолын үндэс
- 3 Микросервис архитектур
- 4 Асуудал ба шийдэл
- 5 Хэрэгжүүлэлт
- 6 Дүгнэлт

Гол зорилго

Хиймэл оюун агентуудыг микросервис архитектурт нэвтрүүлэх боломжийг судалж, **Event-Driven Architecture** ашиглан уян хатан систем бүтээх

Гол ойлголтууд:

- Хиймэл оюуны инженерчлэл
- Микросервис архитектур
- Event-Driven Architecture

Технологи:

- Apache Kafka
- Apache Flink
- RAG систем

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

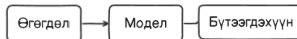
Уламжлалт ML

- Модел хөгжүүлэх
- Өгөгдөл цуглуулах
- Математик мэдлэг

AI Engineering

- Бэлэн суурь модел
- Программ хангамжид интеграц
- Prompt engineering

Машин сургалтын инженерчлэл:



Хиймэл оюуны инженерчлэл:



Хөгжлийн замнал

Хэл модел → Том хэлний модел (LLM) → Суурь модел

Гол үйл явдлууд:

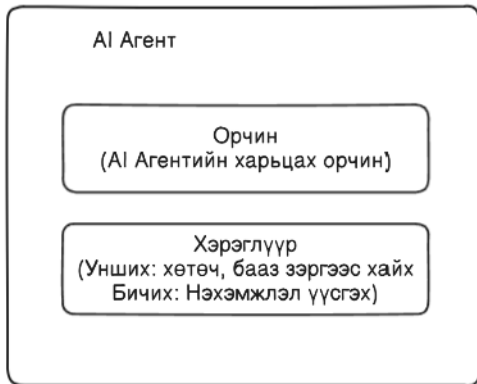
- Transformer (2017)
- Өөрийгөө удирдсан сургалт
- GPT-3: 175 тэрбум параметр
- GPT-4: 1.2 их наяд параметр

Технологи:

- Attention механизм
- Урьдчилан сургалт
- Дараах сургалт
- Sampling стратеги

Агент гэж юу вэ?

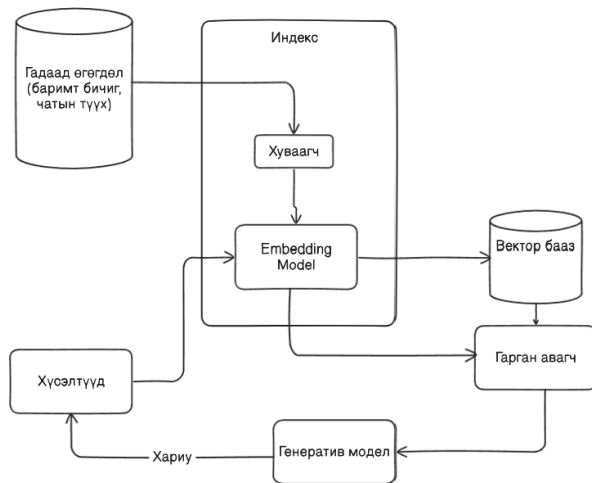
Өөрийн орчныг мэдрэх, түүн дээр үйлдэл хийх чадвартай систем



Гол бүрэлдэхүүн:

- Орчин
- Үйлдэл

RAG (Retrieval-Augmented-Generation)



Монолит

- Нэг том систем
- Нэг кодын сан
- Өргөжүүлэх хүндрэлтэй

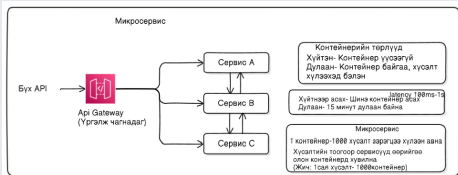
Микросервис

- Жижиг, бие даасан
- Тусдаа хөгжүүлэлт
- Бие даан өргөжүүлэх

Сорилт

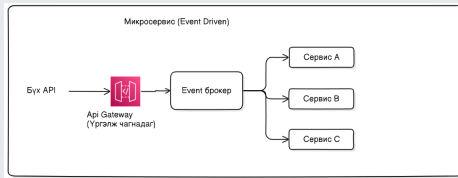
Сервис хоорондын харилцаа, өгөгдлийн нэгтгэл, динамик routing

Синхрон



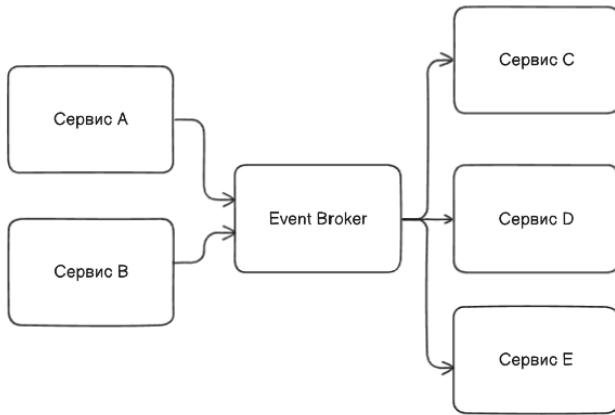
- HTTP REST / gRPC
- NxM нягт холбоос

Асинхрон



- Мессежийн брокер
- Салангид байдал

Event-Driven Архитектур



Үндсэн ойлголтууд

Topic Үйл явдлуудыг ангилах суваг

Producer Үйл явдал үүсгэгч

Consumer Үйл явдал хүлээн авагч

Partition Өргөжүүлэх, параллель боловсруулалт

Давуу талууд

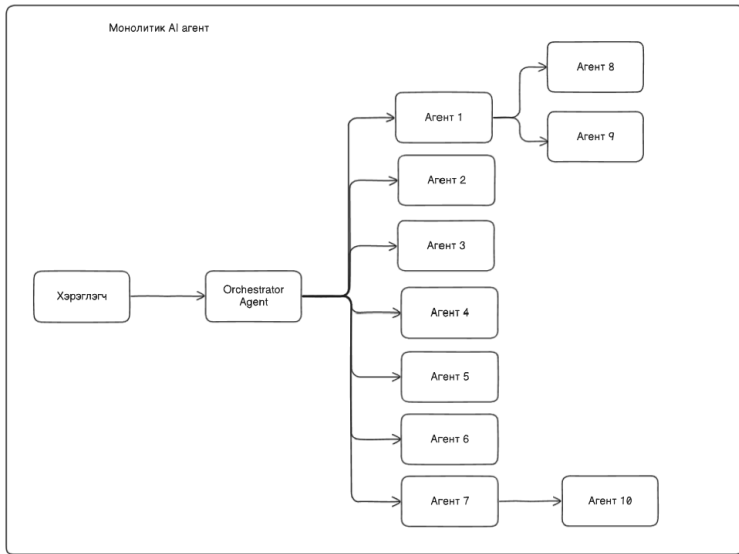
Хэвтээ өргөжих • Бага хоцрогдол • Үйл явдлын хадгалалт • Дахин тоглуулах

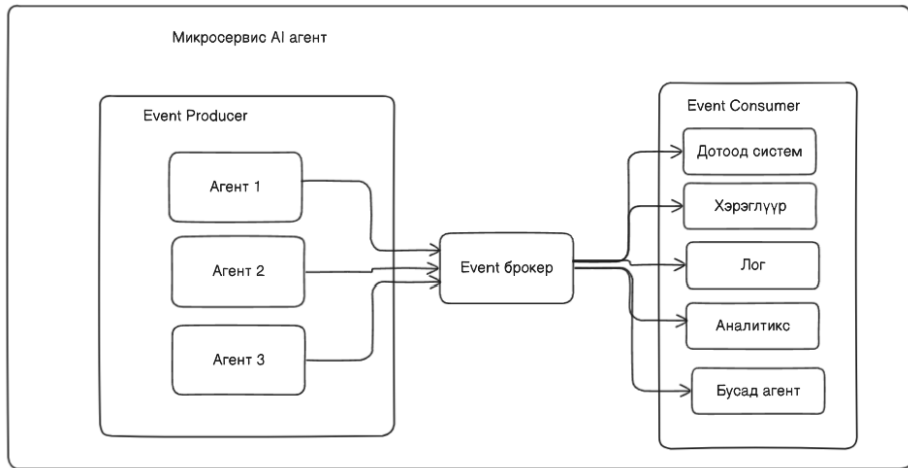
Flink-ийн давуу талууд

- **Төлөв байдлын удирдлага** - Найдвартай 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 Flink app болгон хөгжүүлэх боломжийг олгодог.

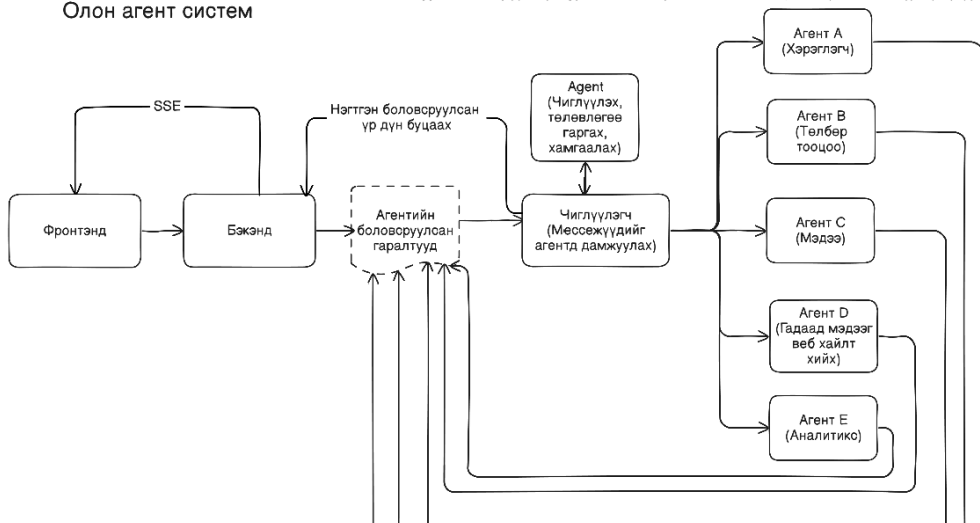




Гол бүрэлдэхүүн хэсгүүд

Олон агент систем

- Агент болгон нь өөрсдийн цар хүрээ дотор даалгавар биелүүлэх үүрэгтэй
- Агентууд бие бие рүүгээ рекурсив байдлаар мэдээлэл солилцоод хэрэглэгчдэд үр дүн үзүүлнэ

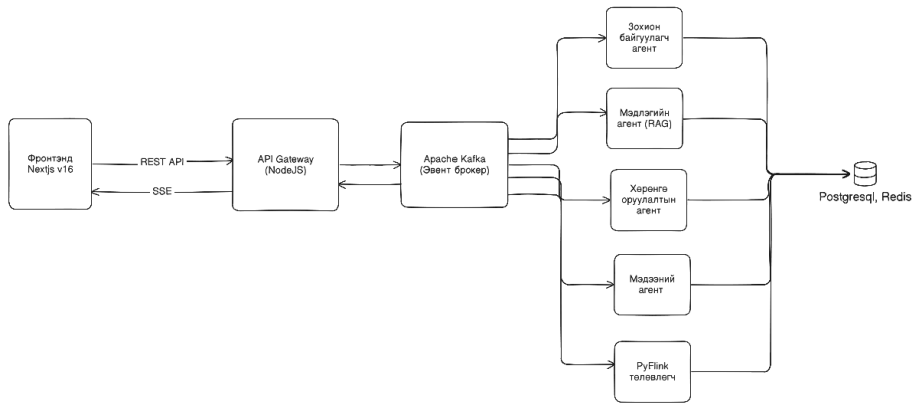


Техникийн давуу талууд

- **Салангид байдал** - Агентууд бие биенээсээ хараат бус
- **Асинхрон харилцаа** - Параллель боловсруулалт
- **Бие даан өргөжүүлэх** - Агент бүр өөрийн хэрэгцээний дагуу
- **Найдвартай байдал** - Үйл явдлууд хадгалагдана

Бизнесийн давуу талууд

Бодит цагийн шийдвэр гаргалт • Лог хадгалалт • Хэвтээгээр өргөжүүлэх • Шинэ агент нэмэхэд хялбар



Системийн бүрэлдэхүүн

Агент	Үүрэг
Orchestrator	Gemini AI ашиглан хүсэлтүүдийг ангиллаж, агентууд руу даалгавар шилжүүлэх
Knowledge	RAG систем, vector search ашиглан мэдлэг хайх
Investment	МХБ өгөгдөл дүн шинжилж, зөвлөмж өгөх
News	Finnhub API-ээр мэдээ цуглуулж хураангуйлах
PyFlink Planner	Рекурсив эсвэл жирийн чиглүүлэг хийх

Frontend

- Next.js 16
- React 19
- TypeScript 5

Backend

- Node.js 20
- Express.js
- PostgreSQL 16
- Redis 7

AI & EDA

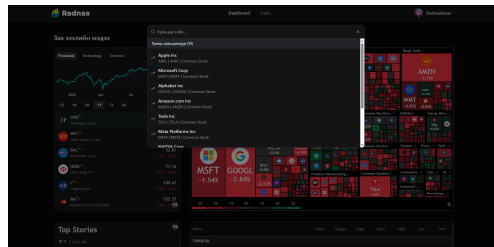
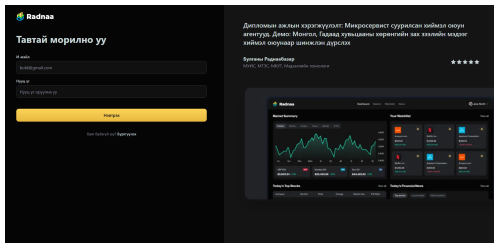
- Google Gemini 2.0/2.5
- Apache Kafka 3.5
- Apache Flink

1. Шинэ хэрэглэгч бүртгүүлэх

Бүртгэл үүсгэх → Профайл оруулах → Kafka event → Gemini AI э-мэйл илгээх

2. Хувьцааны дүн шинжилгээ

Frontend → API Gateway → Kafka → Orchestrator → Investment Agent → SSE streaming



Хөгжүүлэлт

- **Хугацаа:** 6 долоо хоног
- **Хэрэгжүүлэлт:** 70% бүрэн

Бүрэн хэрэгжсэн

Docker, Kafka, PostgreSQL, Redis • Бүх агентууд • API Gateway • Frontend (Vercel)

Гүйцэтгэл

Database: 50-100ms • Kafka: 5-10ms • API Gateway: 200-500ms • AI дүн шинжилгээ: 10-17 сек

Онолын хувьд

- Хиймэл оюуны инженерчлэл нь програм хангамж хөгжүүлэлтийн шинэ салбар
- Суурь модел гарч ирэх нь аппликейшн хөгжүүлэлтийн саад бэрхшээлийг эрс багасгасан
- RAG систем нь агентуудын мэдлэгийг өргөтгөж, найдвартай хариулт өгөх боломжийг олгодог

Практикийн хувьд

- Монолит болон API суурилсан холболт нь NxM нягт хамаарал үүсгэдэг
- EDA ашиглан агентуудыг салангид микросервис болгон хөгжүүлэх шаардлагатай

Техникийн хувь нэмэр

- Хиймэл оюун агентуудыг микросервис архитектурт EDA байдлаар нэвтрүүлэх онол болон практикийн арга зам тодорхойлсон
- Бодит систем хэрэгжүүлснээр түүний үр ашигтай байдлыг харуулсан

Практик хувь нэмэр

МХБ-ийн бодит өгөгдөл дээр суурилсан демо систем • Production орчинд нэвтрүүлж болох технологийн шийдэл

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

Энэхүү судалгааны ажил нь хиймэл оюун агентууд болон микросервис архитектурын уялдааг судалж, **Event-Driven Architecture** ашиглан уян хатан, өргөжих боломжтой систем бүтээх боломжтой гэдгийг онол болон практикийн хувьд харуулсан.

Санал болгосон зохиомж нь цаашид хөгжүүлж, production орчинд нэвтрүүлэх суурь болох бөгөөд хиймэл оюуны технологийг микросервис архитектурт нэвтрүүлэх чиглэлд ач холбогдолтой хувь нэмэр болно.

Баярлалаа!

Асуулт?

Баярлалаа!