Virtual Company System — Deep Analysis & Improvements

ALAI Virtual Company System — Deep Analysis & Improvements

Date: 2026-03-21 **Team:** Petter Graff (Architect), Chip Huyen (ML/RAG), Devil's Advocate (BA)
For: Alem (CEO)

Executive Summary

Sistem ima **solidnu osnovu** ali većina infrastrukture je **neiskorištena ili nefunkcionalna**:

- **16 kompanija postoji** — samo **4 zapravo primaju taskove** (CodeCraft, FlowForge, Lexicon, Resolver)
- **RAG pipeline postoji** (25K+ knowledge chunks) — ali **NIJE integriran** u autonomno izvršavanje
- **Blueprint sistem postoji** — ali ima **1 pokrenuti run koji je failovao**
- **Cross-company bus** — nikad kreirao **nijedan task** (176 logova, 0 matcheva)
- **Company tier_overrides** — definirani u configu ali **potpuno ignorirani** u kodu

Prava vrijednost sistema je **CLAUDE.md injection** — kad pi-orchestrator ubaci company-specific instrukcije u prompt. Sve ostalo je scaffolding koji čeka aktivaciju.

Trenutni Task Flow

```
sequenceDiagram
    participant MC as Mission Control<br/>(5,300 tasks)
```

```
    participant PI as pi-orchestrator<br/>(daemon, 30s poll)
    participant CL as Classifier<br/>(llama3.1:8b)
    participant RT as Router<br/>(HARDCODED map!)
    participant CO as Company<br/>(CLAUDE.md inject)
    participant LLM as Model<br/>(tier 1-5)
    participant HM as HiveMind<br/>(18,974 entries)

    MC->>PI: Poll open tasks (max 2 concurrent)
    PI->>CL: Classify: complexity(1-5), domain
    CL-->>PI: {complexity:2, domain:"code"}

    Note over RT: BUG: Uses hardcoded map<br/>domain-to-company.json IGNORED

    PI->>RT: Map domain → company
    RT-->>PI: CodeCraft

    Note over CO: Only injects first 2000 chars<br/>of CLAUDE.md into prompt

    PI->>CO: Load CLAUDE.md context
    PI->>LLM: Prompt (with company context)

    Note over LLM: BUG: No RAG query here!<br/>25K knowledge chunks unused

    LLM-->>PI: Response
    PI->>HM: feedbackToHiveMind() ← OUTPUT works
    PI->>MC: Update task status

    Note over HM: Knowledge STORED but<br/>never RETRIEVED for next task
```

Kriti?ni Nalazi

1. RAG Gap — Knowledge postoji ali se ne koristi (Chip Huyen)

```
graph LR
    subgraph POSTOJI["Postoji (neiskorišteno)"]
```

```
        K["knowledge.db<br/>25,670 chunks<br/>187MB"]
        H["hivemind.db<br/>18,974 entries<br/>99.3% embedded"]
        F["flywheel.db<br/>11,223 cache<br/>0.053 avg hits"]
        R["retrieval-orchestrator.js<br/>7-store RRF fusion"]
    end

    subgraph RADI["Radi"]
        OUT["Output → HiveMind<br/>feedbackToHiveMind()"]
    end

    subgraph NE_RADI["NE RADI"]
        IN["Input ← RAG<br/>processTaskAsync()<br/>NEMA retrieval step"]
    end

    K -.->|"nikad queried"| IN
    H -.->|"nikad queried"| IN
    OUT -->|"piše"| H

    style NE_RADI fill:#ffcdd2
    style RADI fill:#c8e6c9
    style POSTOJI fill:#fff9c4
```

Fix: Dodaj RAG query u `processTaskAsync()` između classification i prompt construction. **2-4 sata posla**, najveći ROI u sistemu.

2. Company Routing — Config fajl se ne ?ita (Petter Graff)

Problem	Detalj	Lokacija
domain-to-company.json **ignorisan**	Orchestrator koristi hardkodiranu mapu	pi-orchestrator.js:554-567
Company tier_overrides **ignorirane**	getCompanyOverride() uvijek vraća null	pi-orchestrator.js:538
ACTIVE_COMPANY env **nikad setovan**	Skill/MCP resolver ne može raditi	spawn pozivi
"text" domain → Lexicon	Svi non-code taskovi idu na Legal	pi-orchestrator.js:545
Blueprint runner **nikad pozvan**	Orchestrator ne koristi blueprinte	shouldCreatePipeline() unused

3. Company Utilization — 9 od 16 nikad primilo task (Devil's Advocate)

```
pie title Task Distribution po Kompanijama (od 1,186 rutiranih)
    "FlowForge" : 543
    "CodeCraft" : 328
    "Lexicon" : 237
    "Skybound" : 36
    "Datavera" : 17
    "Proxima" : 13
    "Vizu" : 11
    "Ostali (9 kompanija)" : 0
```

Činjenice:

- 40% svih završenih taskova uradio **John ručno** (2,139 od 5,300)
- 22.4% taskova ima `pipeline_company` polje uopšte
- Cross-company bus: **176 logova, 0 kreiranih taskova**
- Blueprint system: **1 run, failovao**
- 9 kompanija: **0 taskova ikad**

Model Tier Routing — Šta radi, šta ne radi

```
graph TB
    subgraph RADI_OK["✅ Radi"]
        T1["Tier 1: llama3.1:8b<br/>Classification"]
        T2["Tier 2: qwen2.5-coder:32b<br/>Code tasks"]
        T3["Tier 3: qwen3:32b / deepseek-r1:70b<br/>Complex reasoning"]
        CB["Circuit breaker<br/>(3 failures → 30s backoff)"]
        FB["ANVIL ↔ FORGE fallback"]
    end

    subgraph NE_RADI2["❌ Ne radi"]
        TO["Company tier_overrides<br/>(getCompanyOverride → null)"]
        T4["Tier 4-5: Claude<br/>(offlineMode=true, disabled)"]
        TT["team-of-teams<br/>(minComplexity=6, disabled)"]
```

```
        ST["Routing stats<br/>(in-memory, lost on restart)"]
        KM["Kimi K2.5 dead code<br/>(llama-server, port 8000)"]
    end

    style RADI_OK fill:#e8f5e9
    style NE_RADI2 fill:#ffebee
```

offlineMode=true — Claude API isključen od 2026-03-19 (budget). Complexity 4-5 taskovi silently downgraded na qwen3:32b.

Improvement Plan — Prioritizirano

P0 — Fix odmah (< 1 dan, najve?i ROI)

#	Fix	Effort	Impact
I1	RAG injection u pi-orchestrator processTaskAsync()	2-4h	**Aktivira 44K knowledge entries**
I2	Učitaj domain-to-company.json umjesto hardcoded mape	30min	Config postaje funkcionalan
I3	Fix getCompanyOverride() da vrati tier_overrides	2-3h	Company model tuning radi
I4	Set ACTIVE_COMPANY env pri spawnu agenta	1h	Skill/MCP resolver radi
I5	Fix "text" → Lexicon default routing	1-2h	Non-code taskovi ispravno rutirani

P1 — Sedmica rada (visoki ROI)

#	Fix	Effort	Impact
I6	Wire blueprint-runner u orchestrator za code taskove	2 dana	ZAKON #18 enforced automatski
I7	Review-cycle feedback loop u cross-company bus	1 dan	Automatski Proveo→CodeCraft fix
I8	Persist routing stats u SQLite	4h	Grafana visibility

#	Fix	Effort	Impact
I9	Re-enable staleTaskCleanup sa heartbeat	4-6h	Stuck tasks auto-cleaned

P2 — Arhitekturalna odluka (CEO)

Odluka	Opcije
Collapse kompanije?	A) Zadrži svih 16 (scaffolding za rast) B) Collapse na 4 aktivne (CodeCraft, FlowForge, Lexicon, Resolver) C) Arhiviraj 9 mrtvih, zadrži 7
Blueprint sistem?	A) Pokreni 1 uspješan E2E run pa proširi B) Arhiviraj kao future capability
Cross-company bus?	A) Fix routing rules da nešto matcha B) Deaktiviraj do kad bude trebao
Claude API budget?	offlineMode=true od 19.03. — C4/C5 taskovi na qwen3:32b. Prihvatljivo?

Kona?na Arhitektura — Šta zapravo radi vrijednost

```
graph TB
    subgraph VALUE["□ Gdje je PRAVA vrijednost"]
        V1["CLAUDE.md injection<br/>Company context u promptu"]
        V2["pi-orchestrator daemon<br/>Auto-routing po domenu"]
        V3["Tier routing<br/>8b → 32b → 70b escalation"]
        V4["HiveMind feedback<br/>Output → knowledge store"]
        V5["Resolver cron<br/>Systemic issue detection"]
    end

    subgraph SCAFFOLDING["□□Scaffolding (postoji, ne radi)"]
        S1["Blueprint phases + gates"]
        S2["96 company skills"]
        S3["Cross-company bus"]
        S4["Company tier_overrides"]
        S5["MCP per-company overlay"]
    end
```

```
    subgraph DEAD["□ Mrtvo"]
        D1["9 kompanija (0 taskova)"]
        D2["Kimi K2.5 pipeline code"]
        D3["team-of-teams (disabled)"]
        D4["alaiml-router-v1 (missing)"]
    end

    style VALUE fill:#c8e6c9
    style SCAFFOLDING fill:#fff9c4
    style DEAD fill:#ffcdd2
```

Preporuka tima

Petter Graff: "Kompanijski layer je skoro potpuno kozmetički na orchestrator nivou. Prioritet: I1 (RAG), I2 (config load), I3 (tier overrides), I6 (blueprint wiring)."

Chip Huyen: "Najveći ROI je RAG injection — 2-4 sata posla, aktivira 44K knowledge entries. Trenutno output loop radi, input loop ne postoji."

Devil's Advocate: "80% vrijednosti postiže se sa 4 kompanije. 9 kompanija ima 0 taskova ikad. Cross-company bus ima 0 kreiranih taskova u historiji. Blueprint ima 1 run koji je failovao."

Expert team review complete. Published to BookStack.

Revision #2
Created 2026-03-21 19:14:06 UTC by John
Updated 2026-05-31 20:05:22 UTC by John