Data Collector ADIA

Intelligent Browser Automation System

CREATED BY

Lamyae Chouinna

Abderahim Ait Smain

Mounir Aithammadi

Amin Aboukir

Charaf Bathahi

Abderhaman Laariqbat

THE CHALLENGE

- **Brittle Automation:** Traditional scrapers break frequently due to CSS changes and dynamic content shifts.
- **Manual Overhead:** Human-in-the-loop data collection is slow, expensive, and unscalable.
- **Complex Anti-Bot:** Modern sites use sophisticated tracking that simple scripts cannot bypass.
- **Scaling Bottlenecks:** Monolithic scrapers struggle with high-concurrency enterprise requirements.

THE ADIA SOLUTION

AI-First Agent

Leveraging LLMs (GPT/Gemini) to reason, navigate, and interact with websites like a human operator.

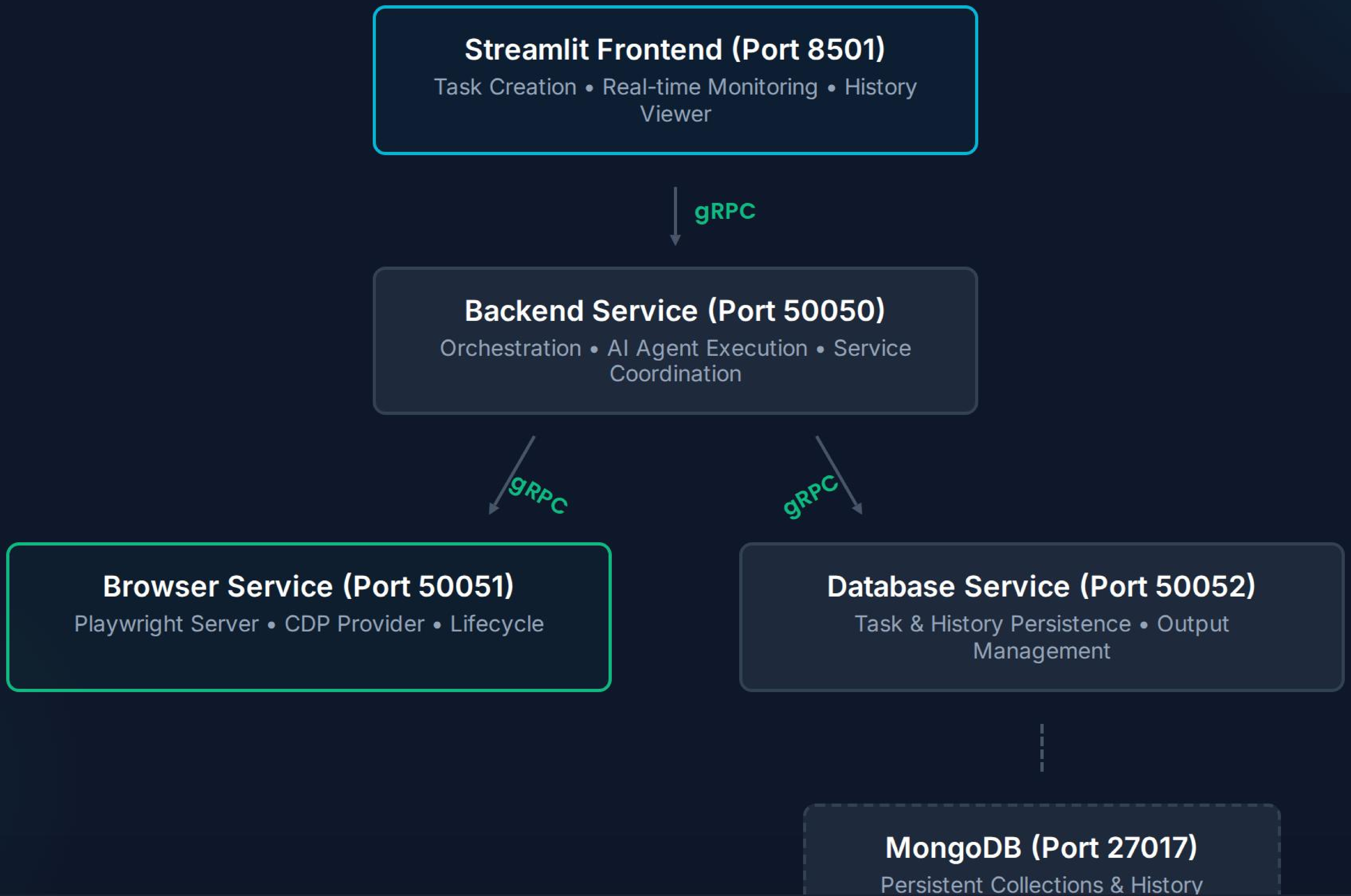
Microservices

Distributed architecture ensuring modularity, horizontal scalability, and high resilience.

Real-time Sync

Task execution status and live browser logs streamed via gRPC and SSE for instant visibility.

SYSTEM ARCHITECTURE



Streamlit Frontend (Port 8501)

Task Creation • Real-time Monitoring • History Viewer

gRPC

Backend Service (Port 50050)

Orchestration • AI Agent Execution • Service Coordination

gRPC

gRPC

Browser Service (Port 50051)

Playwright Server • CDP Provider • Lifecycle

Database Service (Port 50052)

Task & History Persistence • Output Management

MongoDB (Port 27017)

Persistent Collections & History

DISTRIBUTED COMPONENTS

Service Breakdown

Each component operates independently, communicating via type-safe gRPC protocols.

- **Frontend:** Modern Streamlit UI for data scientists.
- **Backend:** The "Brain" coordinating LLM reasoning.
- **Browser:** Isolated Playwright instances for safety.
- **Database:** Scalable MongoDB storage for massive logs.

TECHNOLOGY ECOSYSTEM

Python 3.11+

Core logic utilizing gRPC, FastAPI, and Playwright for high-performance automation.

MongoDB

NoSQL database handling high-velocity task history and semi-structured browser outputs.

Docker

Containerized deployment for seamless environment consistency and cloud readiness.

LLM Integration: OpenAI GPT-4 / Google Gemini **Networking:** Protocol Buffers (Protobuf)

SYSTEM CAPABILITIES

- **Natural Language:** Describe tasks in plain English; the AI handles the selectors.
- **Step-by-Step Monitoring:** Visual feedback of every click and decision the agent makes.
- **Adaptive Execution:** Self-corrects when encountering popups or UI changes.
- **High Performance:** Multi-service design allows concurrent browser sessions.

INTELLIGENT AUTOMATION

The Browser-Use Agent

Utilizing the latest advancements in LLM reasoning to navigate the web autonomously.

- ✓ **Analyze:** Perceives the DOM structure.
- ✓ **Reason:** Decides the next logical step.
- ✓ **Act:** Executes Playwright commands.
- ✓ **Evaluate:** Confirms goal achievement.

EXECUTION WORKFLOW

01

Request

User sends prompt: "Extract top 10 AI news from TechCrunch".

→

02

Spawn

Backend initializes agent and requests browser instance via gRPC.

→

03

Operate

AI Agent navigates, scrolls, and extracts data autonomously.

→

04

Return

Data saved to MongoDB and returned to UI via SSE stream.

INTER-SERVICE COMMUNICATION

Why gRPC?

ADIA leverages gRPC for high-performance internal communication over HTTP/2.

- **Low Latency:** Binary serialization via Protobuf.
- **Type Safety:** Strict contract definition for services.
- **Streaming:** Server-to-client updates for status.

```
// Protobuf Definition
service BrowserService {
  rpc StartBrowser(BrowserRequest) returns
  (CDPResponse);
  rpc StopBrowser(SessionID) returns
  (Status);
}

service DatabaseService {
  rpc CreateTask(TaskData) returns
  (TaskID);
  rpc SaveOutput(OutputData) returns
  (Empty);
}
```

ENGINEERING EXCELLENCE

4

MICROSERVICES

Overcoming Complexity

- **Dynamic Port Allocation:** Automatically handles CDP conflicts for multiple browser instances.
- **DOM Optimization:** Selective element parsing reduces LLM token costs by up to 60%.
- **Retry Resilience:** Built-in exponential backoff for API quota management.
- **Centralized Logging:** Synchronized logs across all 4 services via rotating file handlers.

Thank You

Data Collector ADIA: The Future of Autonomous Web Intelligence

Lamyae Chouinna
Mounir Aithammadi
Charaf Bathahi
Abderahim Ait Smain
Amin Aboukir
Abderhaman Laariqbat

Questions & Discussion

Project Repository: github.com/adia-autonomous/data-collector

IMAGE SOURCES

https://static.vecteezy.com/system/resources/thumbnails/021/050/150/original/abstract-plexus-tech-background-with-glowing-blue-shiny-connecting-lines-and-dots-or-nodes-digital-data-network-connections-concept-this-modern-technology-is-full-hd-and-a-seamless-loop-free-video.jpg

Source: www.vecteezy.com

https://cdn.vectorstock.com/i/1000v/17/49/futuristic-data-dashboard-ui-vector-29231749.jpg

Source: www.vectorstock.com

https://www.siddharthbharath.com/wp-content/uploads/2025/10/Art-Style-Description-Oct-25-2025.png

Source: www.siddharthbharath.com

https://plus.unsplash.com/premium_photo-1764705659986-36d4cea4bf1d?fm=jpg&q=60&w=3000&ixlib=rb-4.1.0&ixid=M3wxMjA3fDB8MHxwaG90by1wYWdlfHx8fGVufDB8fHx8fA%3D%3D

Source: unsplash.com