基于 OpAMP 的可观测性控制平面

及其在 LLM 领域的探索与实践

张晓珣 DaoCloud

Content

目录

Part 01

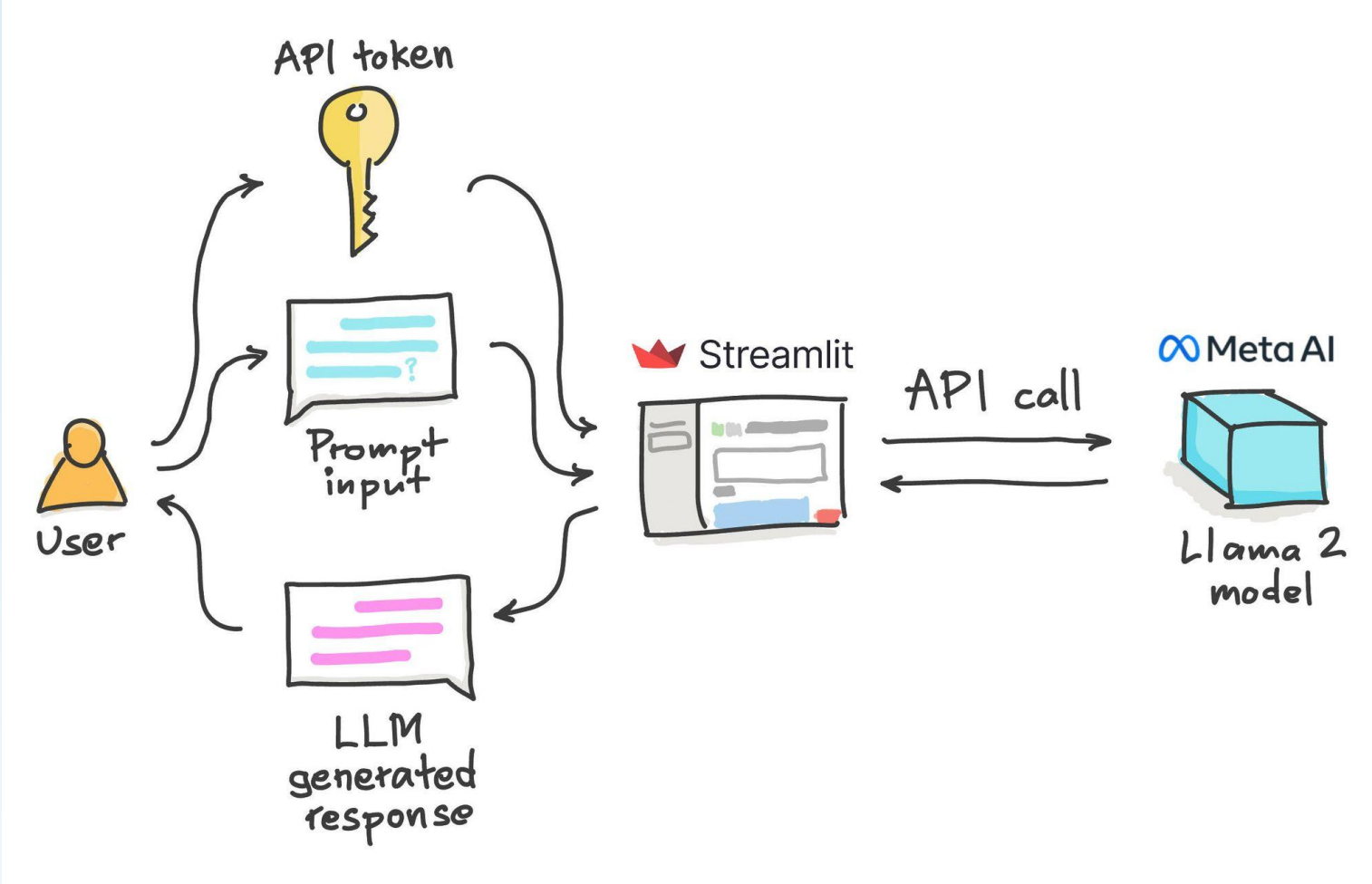
从 LLM 推理应用的观测说起

构建一个 LLM Chatbot

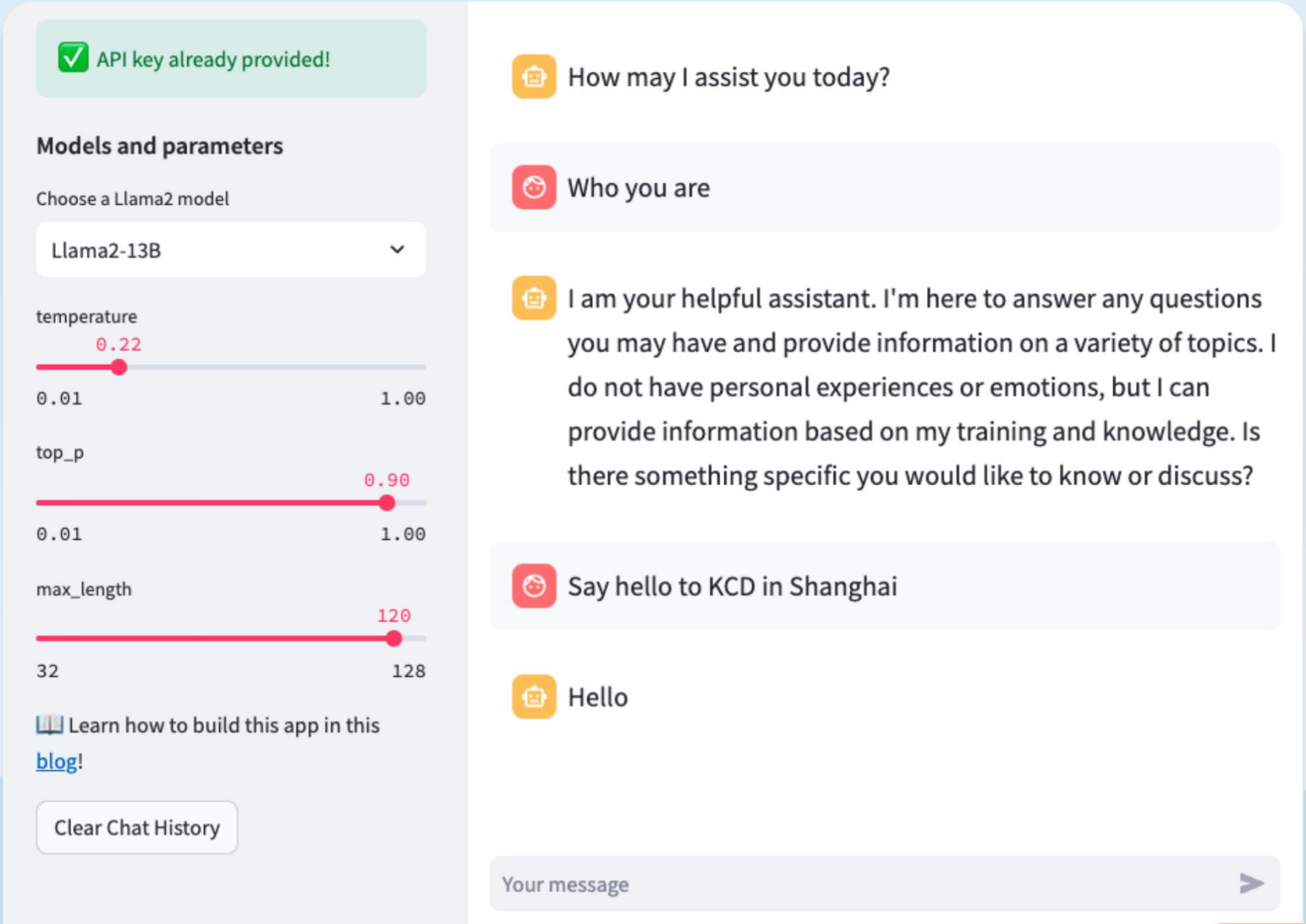
通过亲自创建一个 LLM 聊天机器人应用，来深入学习 LLM 的相关知识。

技术清单：

- 使用 Llama 2 作为 LLM Model
- 利用 Streamlit 构建 UI
- 部署在 Kubernetes 中

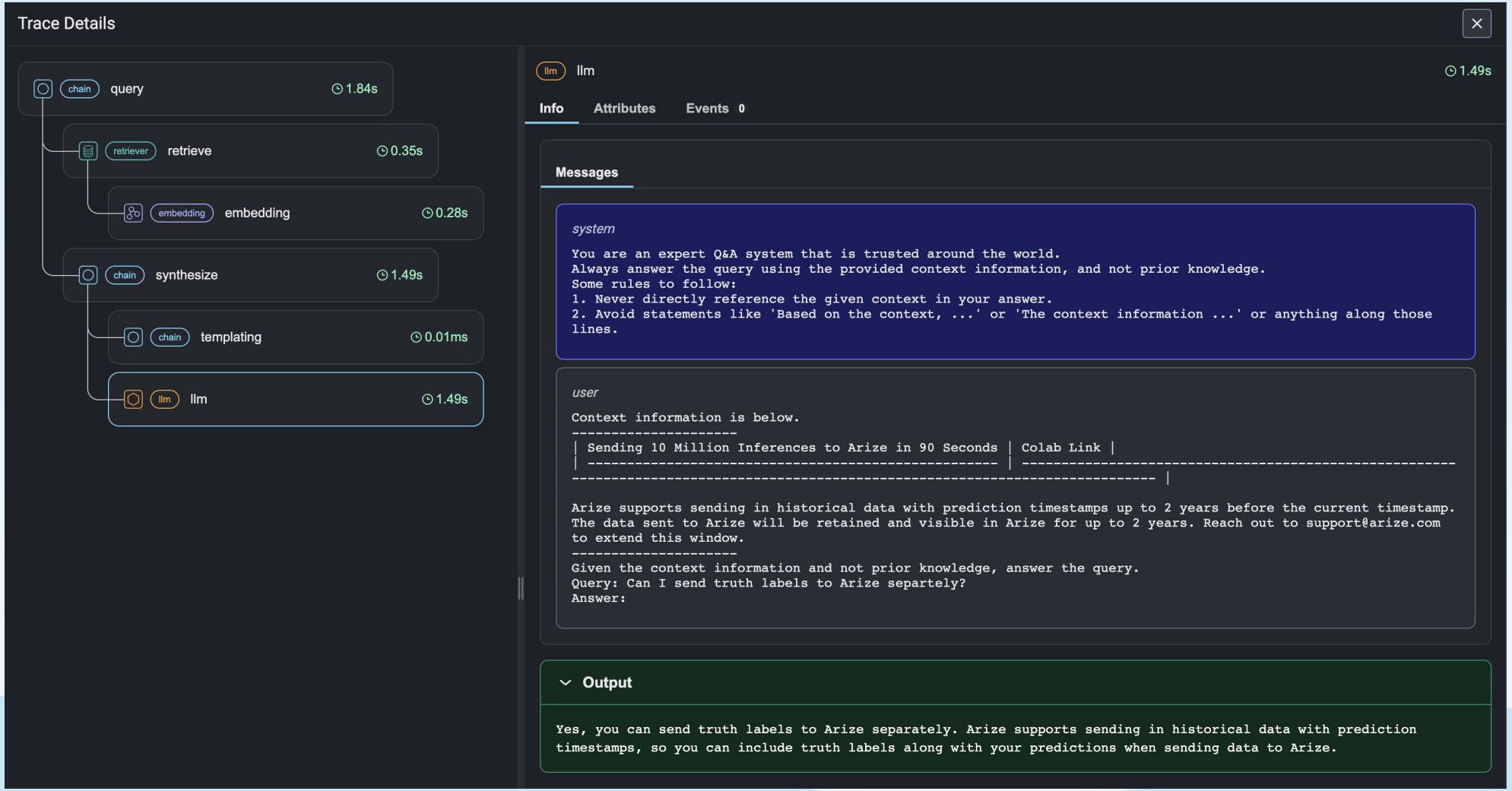


构建一个 LLM Chatbot



LLM 链路

在体验过程中，Bot 的响应速度并不理想。关注到社区的 Arize AI 开源了一个 LLM 的链路工具 Phoenix。



产生了一个疑问？

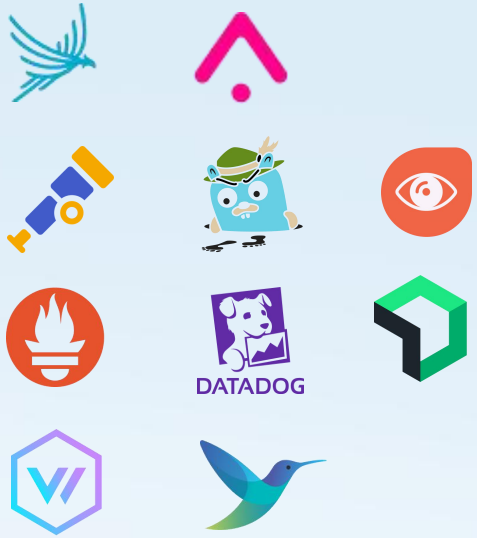
注意到 LLM Observability 社区有一些 LLM 领域专用的观测 方案

LLM Application Layer

AMP Layer

Kubernetes Layer

Infrastructure Layer



进一步思考？

我们需要**更好的方式**来**管理**多样的**采集器**及其配置！

不同维度的观测对象对应的观测方案各不相同。这导致了：

- 采集器组件种类繁多
- 配置繁琐
- 观测需求各异

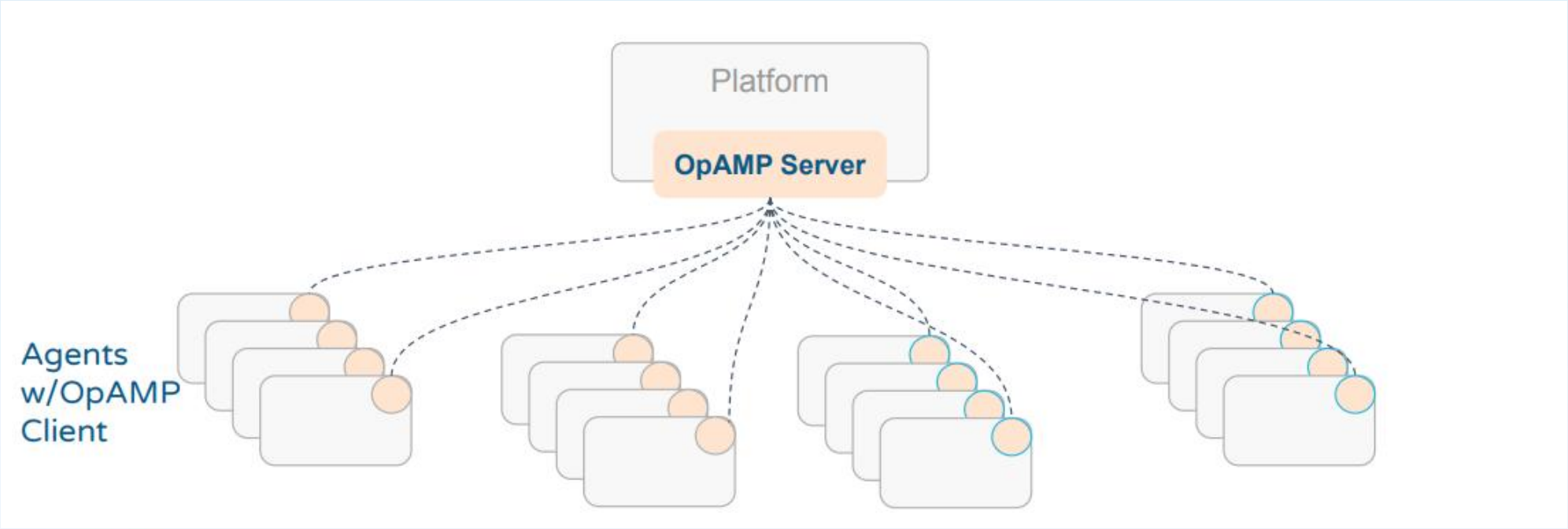
是否引入**可观测性的控制平面**来解决这些问题？

Part 02

OpAMP 构建观测控制平面

Open Agent Management Protocol (OpAMP)

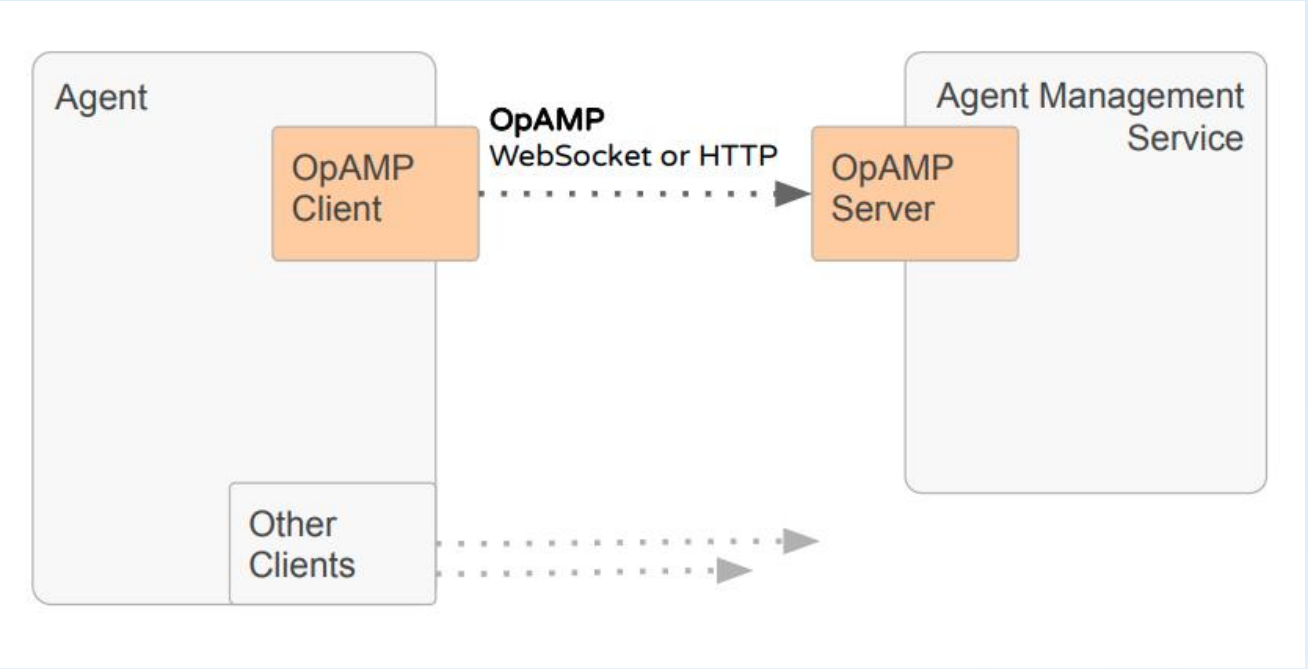
是一种用于远程管理大量数据采集代理(Agents)的网络协议。协议由 OpenTelemetry 组织提出，是供应商中立的，这意味着 OpAMP Server 可以远程管理来自不同供应商的 Agent。



OpAMP 的通信模型 + 接入 Agent

OpAMP 控制平面

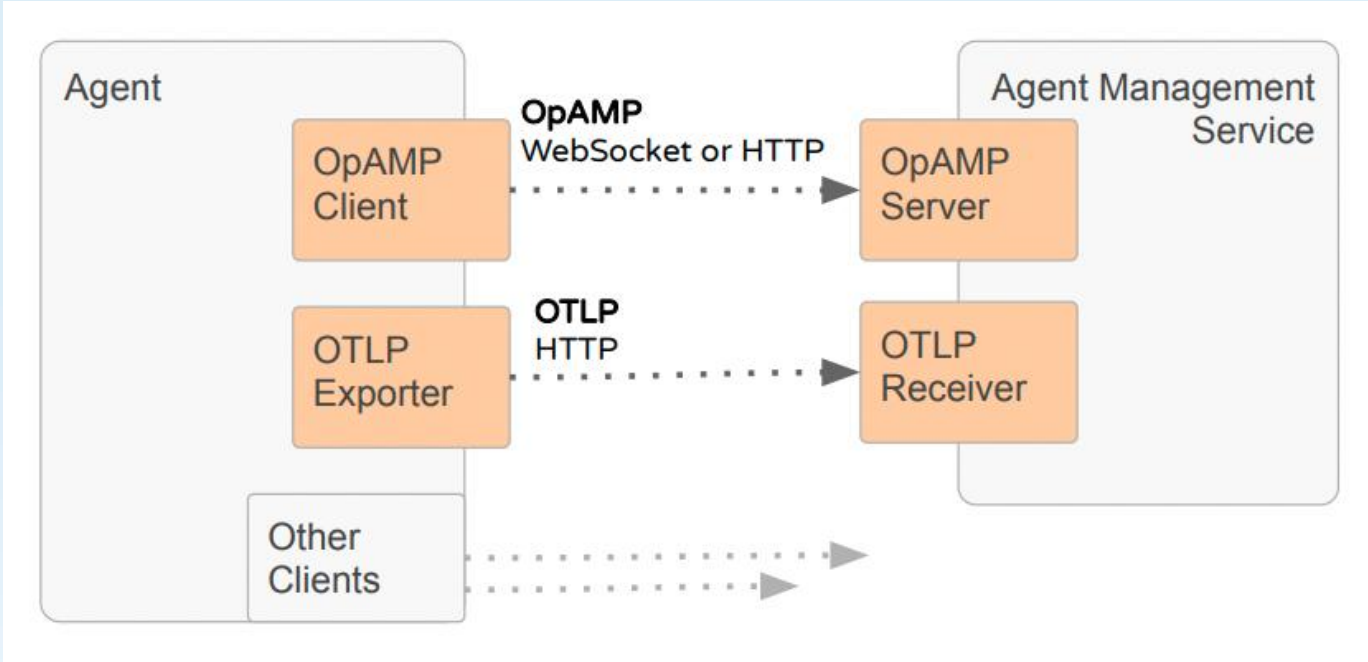
- 区分了 **服务器端** 与 **客户端端**，客户端被客户端纳管，并接收配置
- 在 Agent 中嵌入一个 OpAMP 客户端，**通过 WebSocket 或 HTTP** 与 OpAMP 服务器协议通信
- 传输内容是经过二进制序列化的 Protobuf 报文



OpAMP 的通信模型 + 观测数据平面

观测数据平面

- 让 Agent 对接观测数据后端，通过 OpAMP 配置一个 OTLP exporter 将**观测数据**发送到观测数据后端



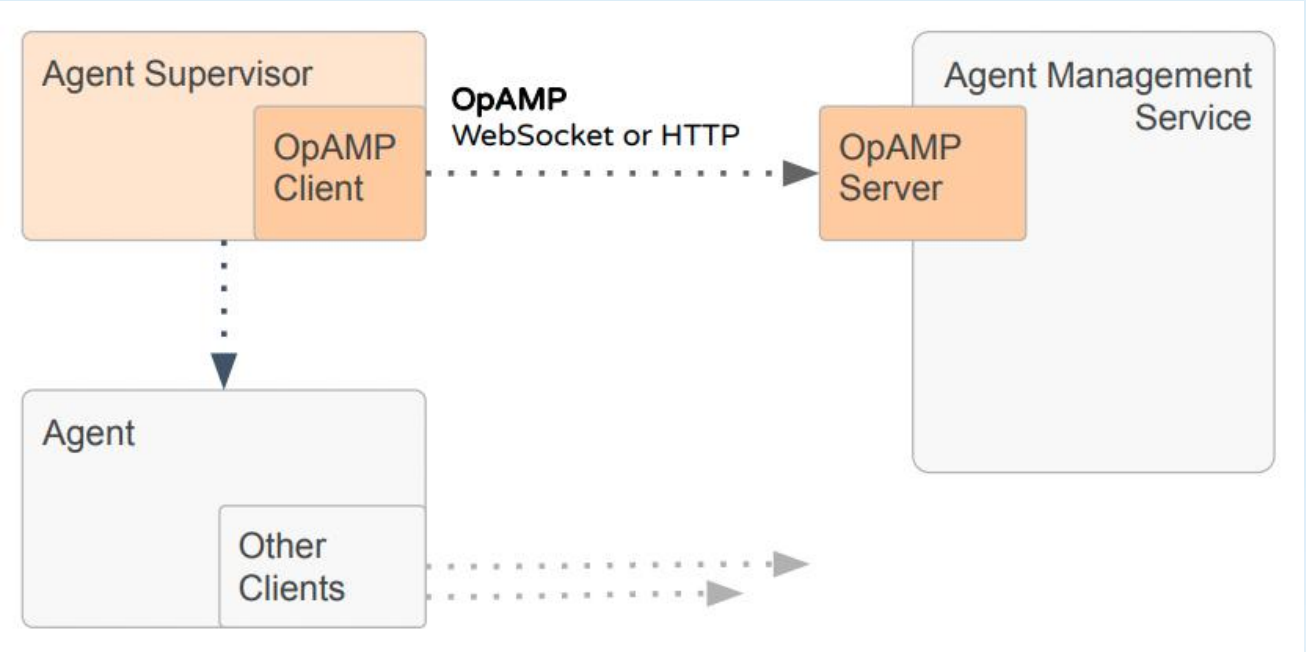
OpAMP 的通信模型 + 监督模式

监督模式 (Supervisor Mode)

- 将 OpAMP 客户嵌入端独立的 **监督员(supervisor) 进程**中，以此管理 Agent

从监督模式延延伸出来的**纳管第三方采集器**：

- 以 Fluent Bit 为例，编写一个专门的**配置转化器**，用于将 OpAMP 服务端下发的配置转化成 Fluent Bit 的配置文件，并**调用 reload 接口**，使配置生效。

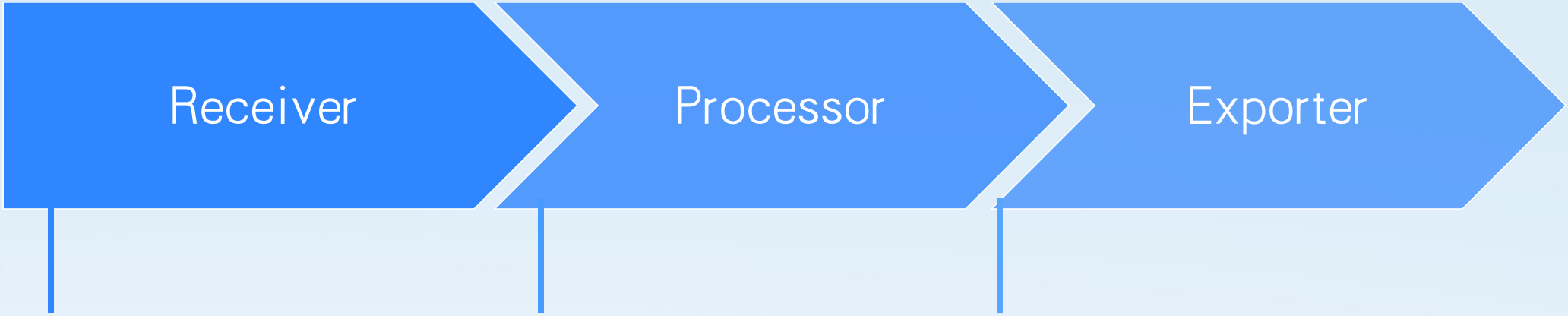


Part 03

OpAMP + OpenTelemetry Collector

OpenTelemetry Collector

penTelemetry Collector



采集接收观测数据

- Prometheus receiver
- HTTP metrics receiver
- OTel receiver
- And over 50+ receivers

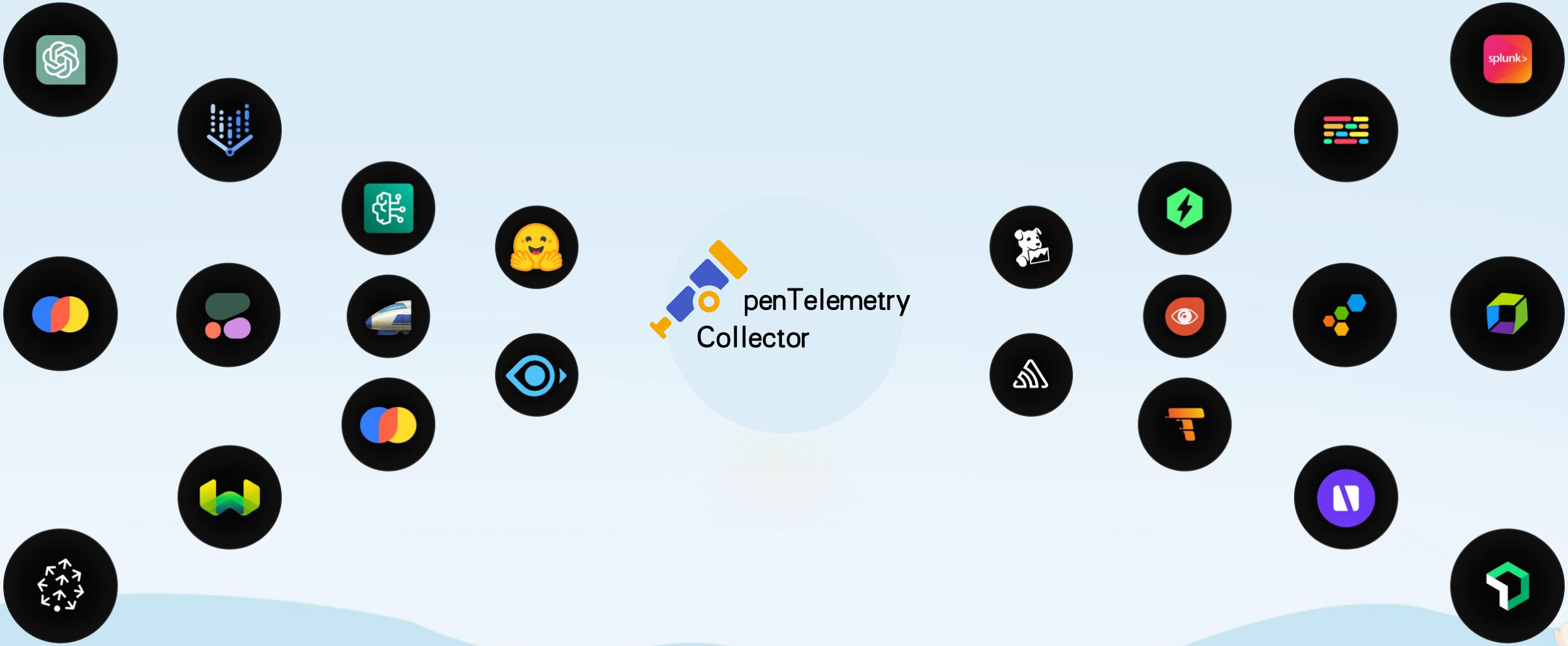
观测数据**处理**

- Kubernetes attributes processor
- Filter processor
- Tail sampling processor

发送数据到观测产品

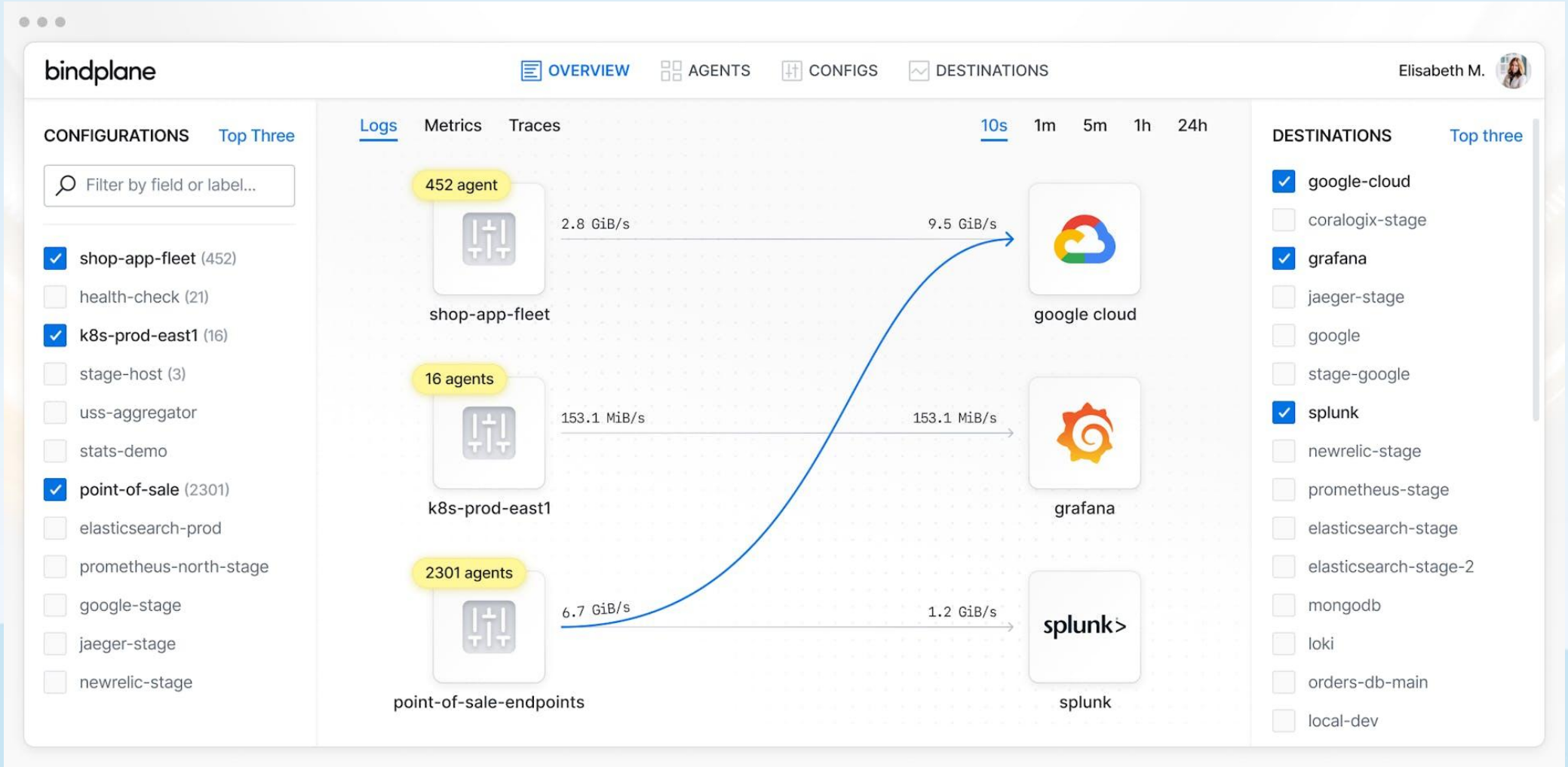
- OTEL

OpenTelemetry Collector 连接一切

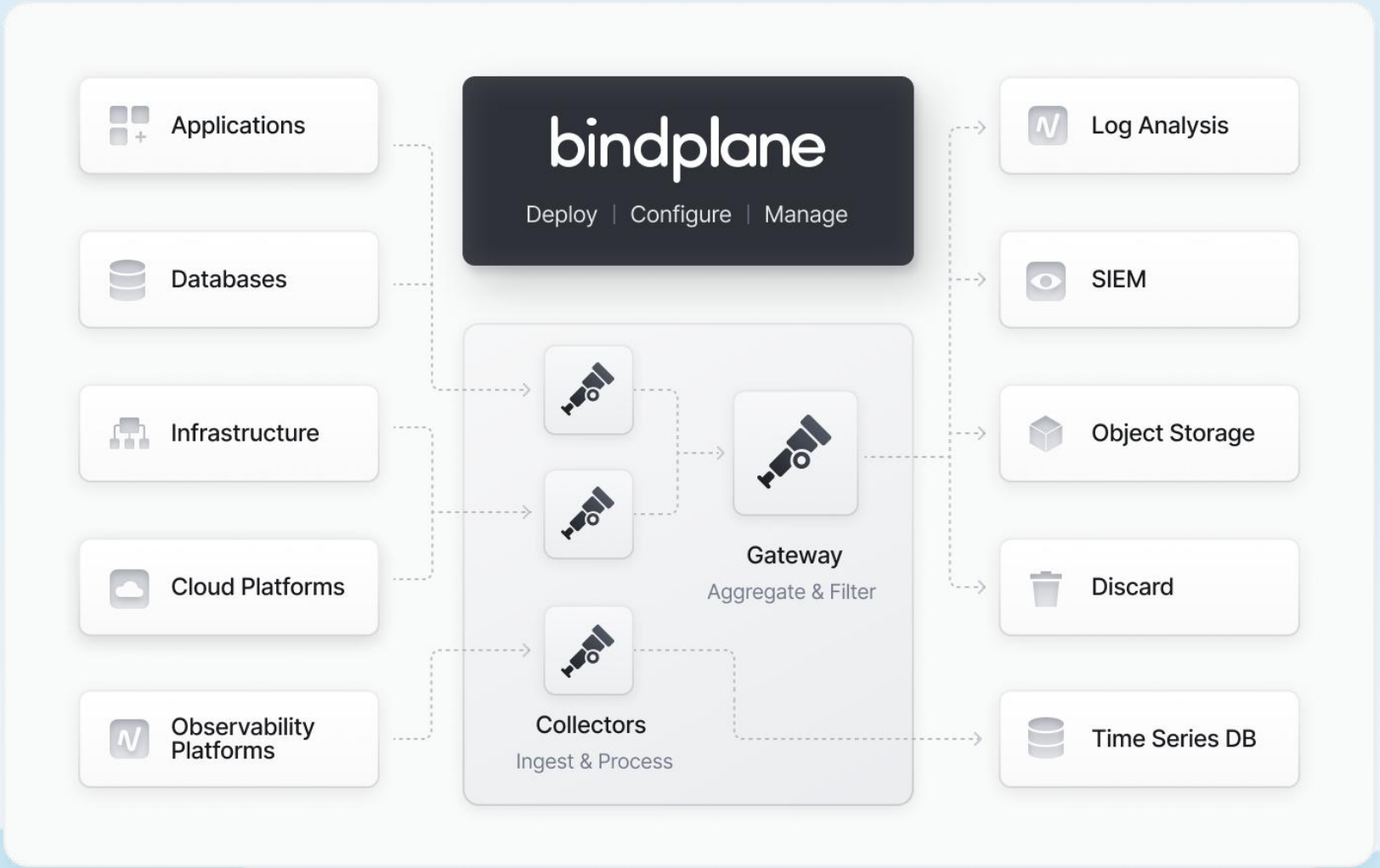


通过 observIQ 的 bindplane

Bindplane 是社区实现 OpAMP 协议的一个工具，通过 OpAMP 和 OpenTelemetry Collector 结合，构建了一个 Observability Pipeline 系统，可以被看作是“观测数据的路由器”，实现了对观测数据的自由分析。



通过 observIQ 的 bindplane



Part 04

再说回 LLM 推理应用的观测

Before

LLM Application Layer

AMP Layer

Kubernetes Layer

Infrastructure Layer

After

LLM Application Layer

AMP Layer

Kubernetes Layer

Infrastructure Layer

penTelemetry
Collector + OpAMP

LLM Observability

DATADOG

Observability Cloud

Open Source

Group: OpAMP &
Opennavigator

欢迎一起交流
云原生观测技术

Thanks.