LEARNING GUIDE / 视频配套资料

Palantir AIP 架构拆解

AI 越强，本体论还要建吗？

沿着一个真实决策，把图读懂。

从 NVIDIA 的物料分配出发，看业务对象、模型、优化器、权限与评估如何配合。

官方原图 / 中文导读 / 访谈重点 / 实践工作表

建议阅读顺序

先看架构全图（第 2 页），再读机制与案例（第 3-4 页）；对照访谈（第 5 页），最后完成一张决策卡（第 6 页）。

我的 Agent 助理

扫描右侧二维码进入。

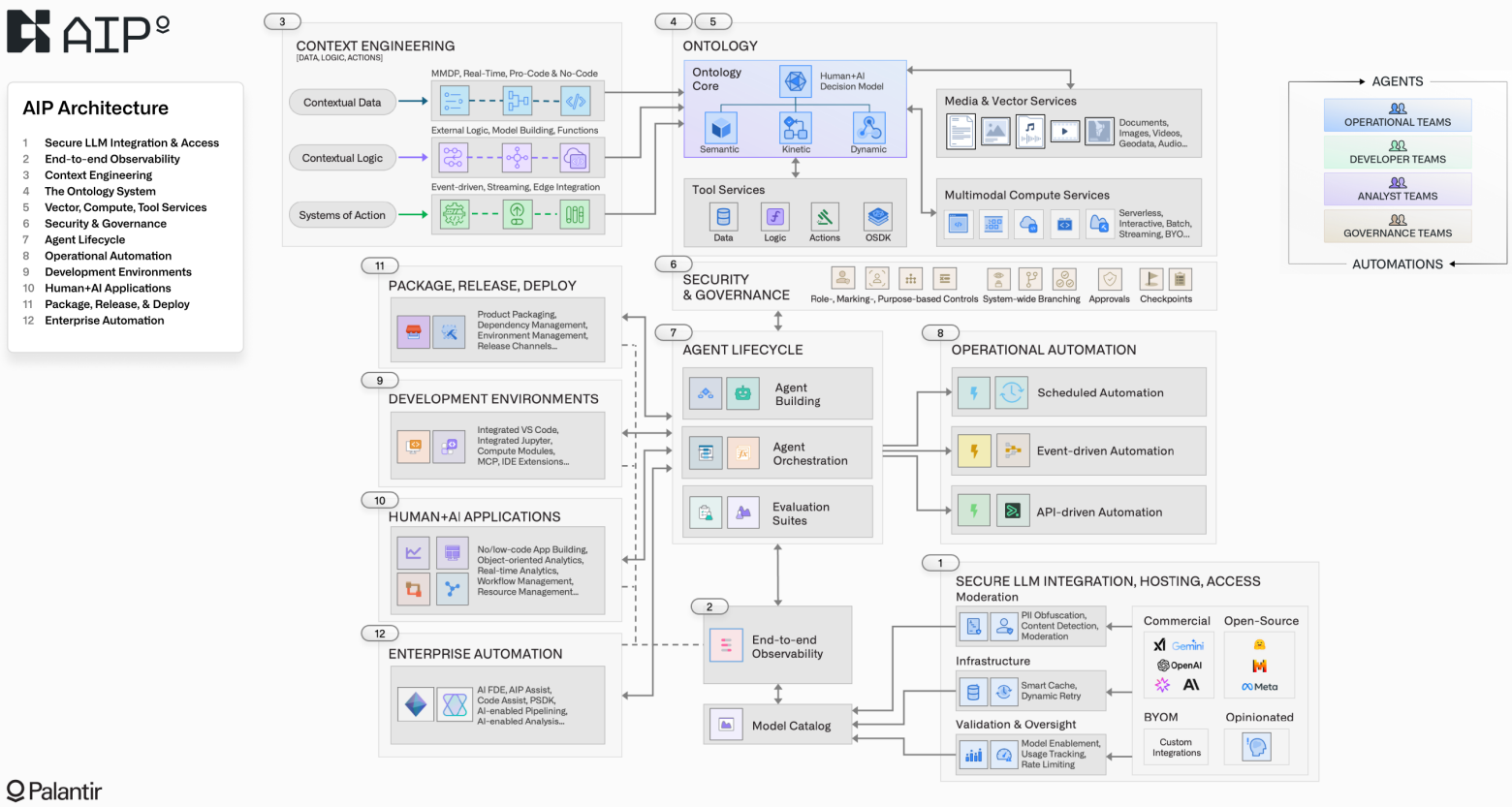
本资料由心晴九天整理。官方事实均附出处；实践判断与练习建议单独标注。

01 / 官方原图

先看连接，再看组件

这是一张能力地图；图中的 1-12 是分类编号，并非执行顺序。

Palantir AIP Architecture Overview



查看高清原图　　阅读官方配套说明

A　业务信息如何进入系统？

左上角的上下文工程连接数据、业务逻辑和可执行操作；中间的 Ontology 将它们组织为共享的业务定义。

B　决策如何执行、如何改进？

向下看 Agent 与自动化，再看安全治理、运行观测、评估和发布。读图时把“处理业务”和“更新系统”两条路径都追完。

阅读提示 | “两条路径”是本资料的读图方法。原图与产品分类来自 Palantir；并非每个项目都需要图中全部能力。

02 / 机制拆解

四个需要分清的边界

弄清每一层承担什么，才能判断模型增强后，哪些工程仍然需要做。

01 业务对象与原始数据

Palantir 的 Ontology 把数据、逻辑、动作与安全组织在一起。学习时既要问“这个对象代表什么”，也要问“谁可以怎样改变它”。

原始资料 02

02 读取权限与提交动作

Action 可以结合用户、参数和对象等信息设置提交条件；能否执行动作，与能否修改动作定义是不同的权限问题。

原始资料 05

03 运行记录与验收标准

AIP Evals 支持定义用例和评价函数、比较版本与模型，并检查多次运行的差异。我的建议：把信息缺失、越权和执行失败也纳入验收。

原始资料 06

04 建设助手与生产变更

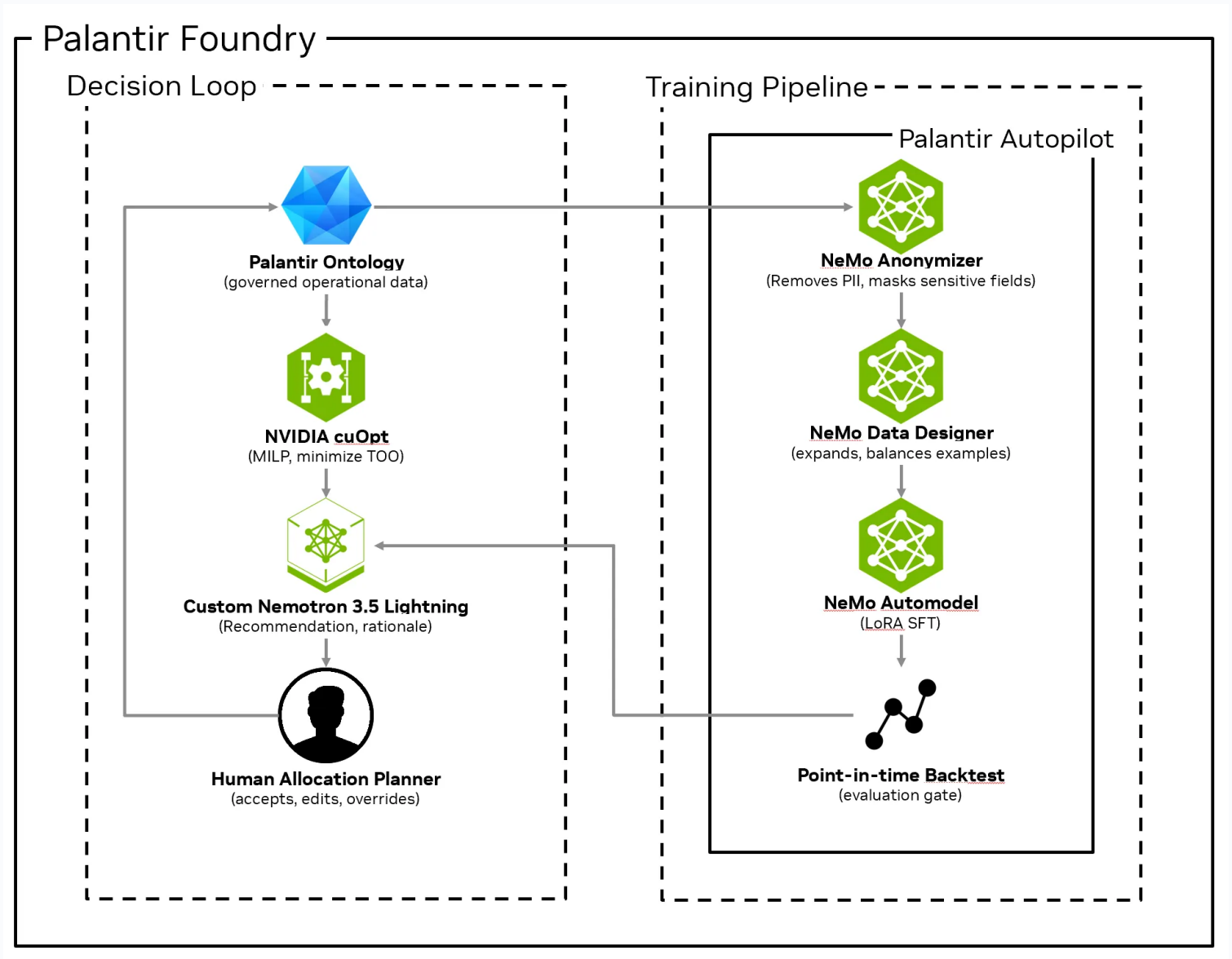
AI FDE 可协助建管道、改 Ontology、写函数；遵循现有用户权限，默认通过分支提案或代码 PR 提交改动供审查。

原始资料 07

03 / NVIDIA 原始案例

有限的物料，怎样分到不同工厂？

案例来源：NVIDIA 技术博客；下图为其公开决策与训练流程图。



打开 NVIDIA 原文与原图

Ontology：提供业务上下文

关联物料、制造地点、产能、分配与实际结果。

cuOpt 与 Nemotron：不同分工

优化器求解数量与产能约束；领域模型结合专家经验给出建议及理由，计划员做最终决定。

历史回放：守住信息时间边界

评估只给决策当天已知的信息，隐藏后续结果；模型不在生产中自行重训。

04 / 访谈导读

带着三个问题看原访谈

Sippin' Sovereignty | A Technical Deep-Dive on Sovereign AI

Chad Wahlquist 与 Sid（Siddhant Singh，Palantir AIP 技术负责人）讨论 NVIDIA 项目。它为这期提供实践背景，并非架构图的逐框讲解。

观看完整访谈 · 约 14 分 42 秒

01:50-03:16

先确定要改善哪个决策

团队先了解数据、权限和计划员的日常工作，聚焦物料分配，到第 4 或第 5 天开始搭工作流。

跳转到这一段

03:38-04:06

业务规则为什么成立？

经同意保留与业务专家的沟通，让工程决定能够追溯到业务依据。

跳转到这一段

06:39-09:55

模型换了，怎样知道有没有退步？

把决策记录与结果关联，沉淀验证用例，区分训练与验证，并以具体业务表现检查模型版本。

跳转到这一段

我的学习判断 | 先写出一项需要改进的业务决策，再决定要建哪些对象与动作。业务定义是否成立，需要业务依据和运行结果共同检验。

05 / 实践工作表

把一个决策写清楚

以下是心晴九天的练习建议，可用来复盘自己的场景。

目标	谁在做什么决策？希望改善哪个结果？
对象	涉及哪些业务对象？唯一标识和关键状态是什么？
依据	哪些事实与规则支持决策？来源、负责人和更新时间在哪里？
分工	哪些步骤需要模型判断？哪些需要规则、计算或优化器？
执行	谁可以提交什么动作？前置条件、写回位置、失败处理是什么？
验证	怎样判断做对了？保留哪些历史用例、异常用例与结果？

我的判断：模型变强之后，保留什么？

模型可能减少字段映射、上下文整理和接口编写的劳动。但多人、多 Agent 共同执行时，业务含义、动作权限与验收标准仍然需要明确，并由系统落实。

一次临时分析未必需要完整本体。是否采用 Palantir，应另行比较业务收益、集成难度与长期维护成本。

06 / 学习入口

原始资料与延伸阅读

点击蓝色标题可打开原文。先读 01-04，再按需要查 05-07。

01 AIP architecture overview

Palantir 官方架构说明与高清原图

02 The Ontology system

数据、逻辑、动作与安全如何共同支持决策

03 Sippin’ Sovereignty

Palantir 官方访谈：Chad 与 Sid 的项目方法

04 NVIDIA 供应链技术案例

From Wafer-Out to First Token：物料分配与模型训练

05 Action submission criteria

业务动作提交时的条件与权限

06 AIP Evals

测试用例、评估标准、模型比较与运行差异

07 AI FDE

能力范围、用户权限以及默认分支审查

我的 Agent 助理

扫描右侧二维码进入。

和我一起继续学习企业 AI、

Ontology 与 Agent 落地。

来源说明 | 图示版权归原作者。中文导读与工作表由心晴九天整理。官方文档及案例核对于 2026.10.11；访谈时间点沿用本期制作时 10.08 的核验记录。