从企业数据到 AI Ready

数据产品与架构实践白皮书

作者待定

目录

前言 7
核心命题 7
本书交付什么 7

白皮书定位与边界 9
目标读者 9
内容边界 9
成功标准 9

阅读与写作指南 11

I 第一部分：为什么企业数据必须走向 AI Ready

1 AI 驱动的组织变化与数据反馈 15
1.1 待回答的核心问题 15
1.2 建议结构 15

2 什么是 AI-Ready Data 17
2.1 工作定义 17
2.2 必须澄清的误区 17
2.3 评价维度 17

3 从基础设施 Ready 到 Agent Ready 19
3.1 四层能力 19
3.2 成熟度原则 19

II 第二部分：把 AI Ready 做成数据产品

4 从 AI 愿景到场景优先级 23
4.1 评估维度 23

5 把 Agent 任务翻译为数据需求 25
5.1 需求分解 25

6 最小可行数据产品与度量 27
6.1 产品最小闭环 27
6.2 指标层次 27

7 数据产品路线图与生命周期 29
7.1 阶段建议 29

III 第三部分：AI Ready 数据参考架构

8 数据源、采集与变化捕获 33
9 处理、存储与数据服务 35
10 上下文层：元数据、语义与契约 37
10.1 核心能力 37
11 语义检索、向量检索与知识关系 39
12 面向 Agent 的数据服务 41

IV 第四部分：可信、治理与持续反馈

13 数据质量、血缘与契约 45
14 权限、隐私与审计 47
15 评估、反馈与持续改进 49

V 第五部分：关键架构决策

16 RAG、语义层还是知识图谱 53
17 批处理、流处理与“及时数据” 55
18 自研、采购与开源组合 57

VI 第六部分：贯穿式企业案例

19 案例：从业务问题到 AI 场景 61
20 案例：数据产品 PRD 63
21 案例：目标架构与关键 ADR 65
22 案例：90 天路线与架构复盘 67

VII 第七部分：从白皮书到组织能力

23 AI Ready Data 的组织运行机制 71
24 从一个场景开始 73
参考文献 75
附录 77
A AI Ready Data 评估表 79
B AI 数据产品 PRD 模板 81
B.1 一、问题与目标 81
B.2 二、数据与上下文 81
B.3 三、产品与 Agent 行为 81
B.4 四、评估与发布 81
C 架构评审清单 83
D 架构决策记录（ADR）模板 85
D.1 ADR-XXX：决策标题 85
D.1.1 背景 85
D.1.2 备选方案 85
D.1.3 决策 85
D.1.4 后果 85
D.1.5 验证与退出 85
E Agent 权限矩阵模板 87
F 高级岗位能力映射 89
参考文献 91

前言

AI 正在改变企业的生产方式：人开始与 Agent 协作，决策和执行周期不断缩短。模型能力只是起点，决定 Agent 能否进入真实业务流程的，是企业能否持续提供及时、准确、可信、可理解、可授权的数据与上下文。

本白皮书讨论的不是“再建一个数据平台”，而是如何把企业既有数据能力升级为可被 AI 稳定消费、可被业务度量、可被组织持续运营的数据产品与架构体系。

核心命题

> AI Ready 不是一项单点技术，也不是把所有数据向量化。它是让数据在明确业务场景下，能够被人、应用和 Agent 安全、可靠、可解释地使用，并形成可评估的反馈闭环。

全书采用四层能力模型：

1. 基**础设**施 **Ready**：计算、存储、网络与弹性资源能够承载 AI 工作负载。
2. 数据工程 **Ready**：数据可采集、可处理、可供给，并满足场景所需的时效。
3. 上下文 **Ready**：元数据、语义、质量、血缘、契约和知识关系让数据可理解、可相信。
4. **Agent Ready**：数据通过 API、检索、工具或 MCP 被安全调用，并具备审计、评估和反馈机制。

本书交付什么

- 一套 AI Ready Data 能力模型与评估表；
- 一套从 AI 场景到数据产品 PRD 的方法；
- 一张企业级参考架构及关键组件职责；
- 一组架构决策、治理策略与实施边界；
- 一个贯穿式企业案例和 90 天落地路线；
- 一套可直接用于评审、汇报和面试的模板工具箱。

白皮书定位与边界

目标读者

- 正在规划企业 AI 数据底座的技术与数据负责人；
- AI 数据产品经理、数据架构师和企业架构师；
- 希望从试点走向规模化落地的数据与 AI 团队。

内容边界

本书以数据产品和数据架构为主线。趋势判断用于解释“为什么现在行动”，但不替代产品方案、架构设计和落地证据。

本书不试图覆盖模型训练、算力芯片或通用大模型算法的全部知识，而聚焦企业数据如何成为 Agent 可持续使用的上下文与行动依据。

成功标准

读者完成阅读后，应能够选择一个 AI 场景，定义它所需的数据产品，画出目标架构，解释关键权衡，制定治理与评估机制，并形成阶段性实施路线。

阅读与写作指南

全书遵循“为什么—做什么—怎么设计—如何取舍—怎样落地”的顺序。

- 第一部分建立问题定义和成熟度模型；
- 第二部分把数据能力转化为可经营的数据产品；
- 第三、四部分分别回答架构和可信治理问题；
- 第五部分集中呈现高级岗位最需要解释的技术权衡；
- 第六部分用一个企业案例串起全流程；
- 第七部分讨论组织运行机制与持续演进；
- 附录提供可复用的评估、PRD、ADR 和评审模板。

每章写作完成时，都要留下一个可验证产物，例如矩阵、模型、架构图、接口契约、指标或决策记录。

I 第一部分：为什么企业数据必须走向 AI Ready

1 AI 驱动的组织变化与数据反馈 15
1.1 待回答的核心问题 15
1.2 建议结构 15

2 什么是 AI-Ready Data 17
2.1 工作定义 17
2.2 必须澄清的误区 17
2.3 评价维度 17

3 从基础设施 Ready 到 Agent Ready . . 19
3.1 四层能力 19
3.2 成熟度原则 19

1. AI 驱动的组织变化与数据反馈

本章产出：组织生产方式、Agent 工作流与数据反馈速度之间的因果链。

1.1 待回答的核心问题

- AI 如何改变知识工作的分工、协作与决策周期？
- 为什么组织提效会提高对数据时效、质量和可解释性的要求?
- 为什么“及时”不是所有数据都实时，而是由业务决策窗口和服务等级决定？

1.2 建议结构

1. 从工业革命到数字化、智能化的生产力演进；
2. 从“人使用系统”到“人与 Agent 共同执行流程”；
3. 数据反馈回路如何限制 Agent 的实际价值；
4. 企业必须建立哪些新的数据供给能力。

2. 什么是 AI-Ready Data

本章产出：可用于全书的统一定义、反例和能力边界。

2.1 工作定义

AI-Ready Data 是能够在明确业务场景下，被模型或 Agent 及时发现、正确理解、可信使用、安全调用并持续评估的数据及其上下文。

2.2 必须澄清的误区

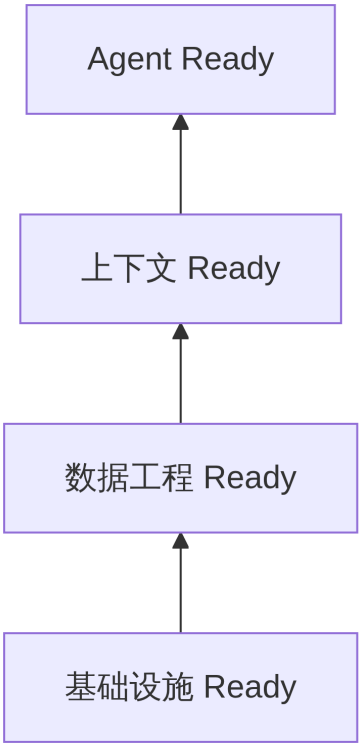
- AI Ready 不等于把数据全部放进向量数据库；
- AI Ready 不等于数据仓库更名；
- AI Ready 不等于所有链路实时化；
- AI Ready 不等于只追求模型能够读取；
- AI Ready 是场景相关的，不存在脱离用途的绝对 Ready。

2.3 评价维度

可发现、可访问、可理解、可信、及时、安全、可操作、可评估。

3. 从基础设施 Ready 到 Agent Ready

本章产出：四层成熟度模型和各层进入条件。



3.1 四层能力

1. 基础设施层：算力、存储、网络、资源调度与成本；
2. 数据工程层：采集、加工、建模、批流处理与服务；
3. 上下文层：目录、元数据、语义、质量、血缘、契约与知识；
4. Agent 层：检索、API、工具、MCP、权限、审计、评估与反馈。

3.2 成熟度原则

成熟度不是采购清单。每一层都要用业务场景、服务等级和可观测指标验证，并允许不同数据域处于不同阶段。

II 第二部分：把 AI Ready 做成数据产品

4 从 AI 愿景到场景优先级 23
4.1 评估维度 23

5 把 Agent 任务翻译为数据需求 25
5.1 需求分解 25

6 最小可行数据产品与度量 27
6.1 产品最小闭环 27
6.2 指标层次 27

7 数据产品路线图与生命周期 29
7.1 阶段建议 29

4. 从 AI 愿景到场景优先级

本章产出： AI 场景优先级矩阵。

4.1 评估维度

- 业务价值与影响范围；
- 数据可得性和数据风险；
- 任务可评估性；
- 流程可嵌入性；
- 实施成本与时间；
- 错误后果及人工兜底能力。

优先选择价值明确、数据边界清晰、反馈可收集、失败可控的场景，避免一开始就建设无法验证价值的“大而全”平台。

5. 把 Agent 任务翻译为数据需求

本章产出： Agent 数据需求说明书。

5.1 需求分解

1. Agent 要完成什么任务、做出什么行动？
2. 它需要哪些事实、规则、历史记录和实时状态？
3. 每项数据的允许延迟、准确率和覆盖率是多少？
4. 哪些字段可见、哪些动作需要审批?
5. 如何证明回答来源并重放决策过程？
6. 哪些反馈会进入下一轮数据和模型改进?

最终需求不能只写“接入知识库”，而要落到数据实体、语义、接口、服务等级和验收指标。

6. 最小可行数据产品与度量

本章产出： 最小可行数据产品（MVDP）边界和指标树。

6.1 产品最小闭环

- 有明确消费者：人、应用或 Agent；
- 有稳定输入、处理规则和输出契约；
- 有负责人、服务等级和变更机制；
- 有质量、使用、价值和成本指标；
- 有用户反馈和退出机制。

6.2 指标层次

数据指标不等于业务价值。指标应从数据可用性、Agent 任务质量、流程效率到最终业务结果逐层建立因果关系。

7. 数据产品路线图与生命周期

本章产出： 从试点、产品化到规模化的阶段路线图。

7.1 阶段建议

1. **验证期**：一个场景、少量数据源、人工兜底；
2. **产品化期**：稳定契约、质量门禁、权限和可观测性；
3. 复用期：跨场景共享语义、检索和数据服务；
4. **规模化期**：域自治、平台治理和价值组合管理。

路线图应同时表达业务里程碑、产品能力、架构能力、治理要求和组织职责。

III 第三部分：AI Ready 数据参考架构

8 数据源、采集与变化捕获 33

9 处理、存储与数据服务 35

10 上下文层：元数据、语义与契约 37

10.1 核心能力 37

11 语义检索、向量检索与知识关系 39

12 面向 Agent 的数据服务 41

8. 数据源、采集与变化捕获

本章产出：企业数据源分类和采集策略矩阵。

需要覆盖结构化数据、文档与多媒体、事件流、SaaS API、日志以及人工知识。针对每类数据，比较批量同步、CDC、流式采集、API 拉取和文件交换的适用条件。

关键设计项包括数据所有权、源端影响、模式变化、重复与乱序、敏感数据识别、失败恢复和采集成本。

9. 处理、存储与数据服务

本章产出： 面向 AI 场景的数据处理与服务分层。

本章讨论湖仓、流处理、特征与指标、文档解析、数据缓存和服务接口之间的职责边界。

架构选择应从访问模式倒推：分析查询、低延迟事实查询、上下文检索、批量训练和 Agent 工具调用，对数据形态和服务等级的要求并不相同。

10. 上下文层：元数据、语义与契约

本章产出： Context Layer 概念模型及核心实体关系。

10.1 核心能力

- 技术、业务和运行元数据；
- 数据目录、术语表和领域模型；
- 数据血缘、质量结果和可观测事件；
- 数据所有者、策略、标签与访问上下文；
- 数据契约、版本和变更影响；
- 面向 Agent 的可检索描述和可信度信号。

本章将结合 OpenMetadata 的实体、连接器、血缘、质量、治理和 API 思想，分析如何让机器不仅“找到数据”，还能够理解数据的含义、来源、状态和使用边界。

11. 语义检索、向量检索与知识关系

本章产出：不同检索和知识组织方式的选择矩阵。

关键词检索、向量检索、结构化查询、语义层和知识图谱解决的问题不同。企业方案通常需要组合，而不是寻找单一“万能检索”。

本章重点比较召回范围、准确性、可解释性、更新成本、权限过滤和评估方法，并说明 Chunk、Embedding、索引与源数据版本之间的可追溯关系。

12. 面向 Agent 的数据服务

本章产出： Agent 数据访问层和受控行动链路。

数据可以通过 SQL、语义查询、REST/GraphQL API、检索服务、工具调用或 MCP 暴露给 Agent。选择接口时，需要同时考虑任务抽象、稳定性、最小权限、上下文预算、延迟、审计和幂等性。

“能读”与“能行动”必须分开治理。高风险写操作应设计审批、策略校验、沙箱、回滚和人工接管机制。

IV 第四部分：可信、治理与持续反馈

13 数据质量、血缘与契约 45

14 权限、隐私与审计 47

15 评估、反馈与持续改进 49

13. 数据质量、血缘与契约

本章产出：从数据问题到 Agent 影响的端到端可观测模型。

传统质量规则关注表和字段，AI Ready 还要关注文档解析、Chunk、Embedding、索引、语义定义和回答引用。血缘需要贯通源数据、派生数据、知识表示、检索结果和 Agent 输出。

数据契约用于明确生产者与消费者之间的结构、语义、质量、时效和变更承诺，并通过自动化门禁控制破坏性变化。

14. 权限、隐私与审计

本章产出： Agent 场景的权限矩阵和审计事件模型。

权限设计需要覆盖身份、数据、字段、行、用途、工具和动作。Agent 不能因为代表用户工作，就自动继承用户的全部能力。

本章讨论最小权限、授权代理、动态策略、敏感信息脱敏、提示注入防护、数据泄露控制、操作审批以及从回答到来源、提示、工具和执行结果的审计链。

15. 评估、反馈与持续改进

本章**产**出： 离线评估、线上监控和业务反馈的闭环。

评估应区分数据质量、检索质量、生成质量、工具执行质量和业务结果。单一回答准确率无法解释系统问题出在数据、检索、模型还是流程。

反馈数据本身也是数据产品：需要定义采集时机、标签体系、隐私边界、归因方式和进入下一轮优化的流程。

V 第五部分：关键架构决策

16 RAG、语义层还是知识图谱 53
17 批处理、流处理与“及时数据” 55
18 自研、采购与开源组合 57

16. RAG、语义层还是知识图谱

本章产出：三类方案的决策树和组合架构。

比较维度包括数据形态、问题类型、关系复杂度、计算精度、解释要求、更新频率、治理成本和团队能力。

本章的重点不是宣布胜出者，而是解释何时用结构化查询，何时用 RAG，何时引入语义层或知识图谱，以及它们如何共享统一的元数据和权限上下文。

17. 批处理、流处理与“及时数据”

本章产出：业务决策窗口到技术时效方案的映射表。

实时是一种成本较高的架构承诺。应从业务允许的决策延迟、数据变化速度、错误成本和恢复要求，倒推出批次、微批、CDC、流式或请求时查询。

本章将给出时效服务等级的定义方式，并分析一致性、乱序、重复、回放、状态管理和成本之间的权衡。

18. 自研、采购与开源组合

本章产出： Build/Buy/Open Source 评估框架和 ADR 示例。

评估维度包括差异化价值、集成复杂度、数据主权、扩展性、人才结构、总拥有成本、退出成本和供应商风险。

OpenMetadata 等开源项目既可以作为组件，也可以作为理解现代元数据平台领域模型和工程边界的参考。引用源码时必须区分“产品现状”“设计思想”和“本书建议”。

VI 第六部分：贯穿式企业案例

19 案例：从业务问题到 AI 场景 61

20 案例：数据产品 PRD 63

21 案例：目标架构与关键 ADR 65

22 案例：90 天路线与架构复盘 67

19. 案例：从业务问题到 AI 场景

本章产出：一个贯穿全书的匿名化企业背景、问题树和基线指标。

案例应具有多数据源、知识分散、权限复杂和反馈周期较长等典型特征，同时保留可落地的组织规模。业务目标、用户、流程、数据现状、约束和错误后果必须明确。

个人经历中的真实洞察可以使用，但企业名称、人员和敏感指标应做匿名化或区间化处理。

20. 案例：数据产品 PRD

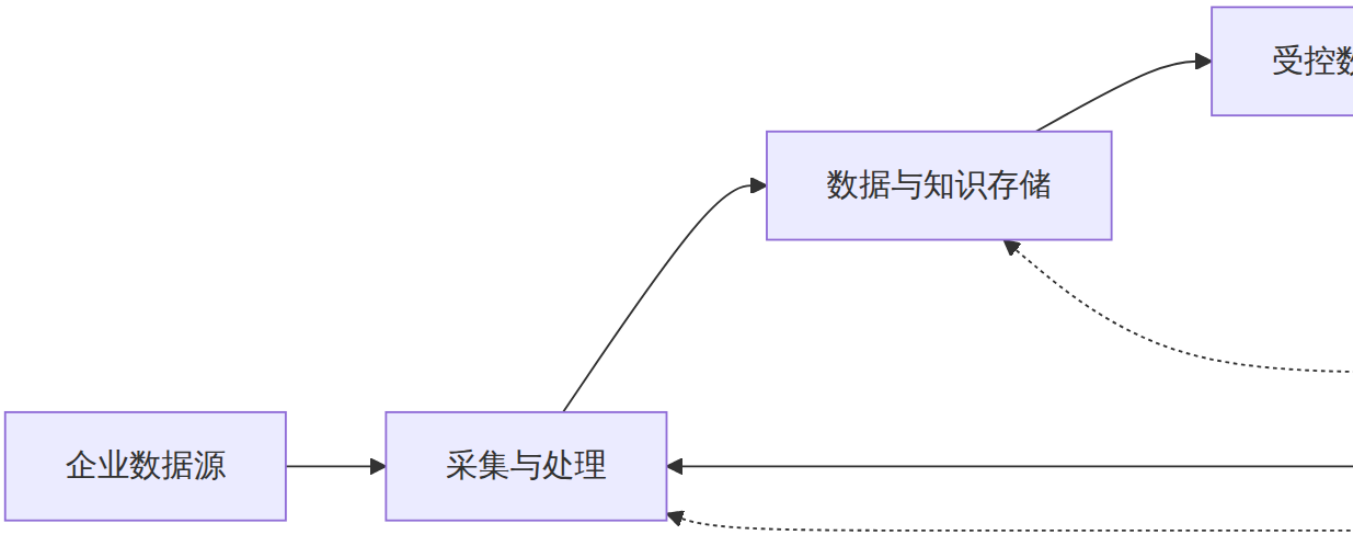
本章产出： 一份完整的 AI 数据产品 PRD。

PRD 至少包括目标用户、待完成任务、场景边界、数据输入、产品输出、服务等级、权限、质量指标、评估集、人工兜底、依赖、风险和验收标准。

需要明确区分用户想要的功能、Agent 完成任务所需的能力以及底层数据平台提供的通用能力。

21. 案例：目标架构与关键 ADR

本章产出： 现状架构、目标架构、数据流和至少三份关键 ADR。



每个组件都要说明职责、输入输出、服务等级、负责人和失败模式，避免只画产品 Logo。

22. 案例：90 天路线与架构复盘

本章产出： 30/60/90 天实施计划、风险清单和复盘框架。

- 0—30 天：场景确认、数据盘点、基线和可行性验证；
- 31—60 天：最小闭环、权限、质量、评估与人工兜底；
- 61—90 天：稳定化、观测、运营机制和下一场景复用。

复盘不仅汇报上线，还要比较目标与结果、分析未达成原因、核算成本与收益，并决定扩展、调整或停止。

VII 第七部分：从白皮书到组织能力

23 AI Ready Data 的组织运行机制 71

24 从一个场景开始 73

24 参考文献 75

23. AI Ready Data 的组织运行机制

本章产出： 产品、平台、数据域、安全和业务团队的职责模型。

AI Ready Data 需要产品化运营，而不是项目交付后结束。数据产品负责人对消费者价值和服务等级负责；数据域对语义与质量负责；平台团队提供通用能力；安全与治理团队提供政策和监督；业务团队参与反馈和价值验证。

本章讨论集中治理与域自治的边界，以及架构评审、数据契约、价值复盘和能力社区如何形成长期机制。

24. 从一个场景开始

AI Ready 是一条从业务结果倒推数据、上下文和架构能力的演进路线，不是一次性的平台重建。
建议从一个有价值、可评估、风险可控的场景开始：

1. 明确任务与业务基线；
2. 评估所需数据的 Ready 程度；
3. 定义最小可行数据产品；
4. 建立受控的数据与 Agent 闭环；
5. 用指标和反馈决定下一步投资。

企业最终需要的不是更多孤立的 AI 演示，而是一套能够反复把数据转化为智能行动的组织能力。

参考文献

附录

A. AI Ready Data 评估表

维度	核心问题	证据	当前等级	目标等级	负责人
可发现	Agent 能否定位正确数据？	目录覆盖率、检索成功率			
可访问	是否有稳定、受控的访问方式？	API/查询可用性			
可理解	是否有业务语义和机器可读上下文？	术语、描述、语义映射			
可信	质量、血缘和新鲜度是否透明?	质量结果、血缘、SLA			
及时	是否满足业务决策窗口？	端到端延迟			
安全	是否遵循最小权限与用途限制？	策略、审计、脱敏			
可操作	是否能支持受控工具调用或行动？	工具契约、审批、回滚			
可评估	是否能度量任务和业务结果？	评估集、反馈、业务指标			

B. AI 数据产品 PRD 模板

B.1 一、问题与目标

- 目标用户：
- 待完成任务：
- 当前流程与基线：
- 目标业务结果：
- 不在范围内：

B.2 二、数据与上下文

- 数据实体和来源：
- 业务语义：
- 时效与质量要求：
- 权限与隐私：
- 来源引用与可追溯：

B.3 三、产品与 Agent 行为

- 输入、输出与工具：
- 正常路径：
- 异常路径和人工兜底：
- 服务等级：

B.4 四、评估与发布

- 离线评估集：
- 线上产品指标：
- 业务指标：
- 灰度、回滚和退出标准：

C. 架构评审清单

- ☐ 业务场景、用户和成功指标清晰；
- ☐ 数据源、负责人和合法用途明确；
- ☐ 时效由业务决策窗口推导；
- ☐ 数据结构、语义、质量和变更契约明确；
- ☐ 权限在检索、生成和行动各阶段一致执行；
- ☐ 血缘贯通源数据、索引、回答和行动；
- ☐ 失败模式、降级、重试、回放和人工接管已设计；
- ☐ 评估能够区分数据、检索、模型和工具问题；
- ☐ 成本、容量、可观测性和运维责任明确；
- ☐ 关键选择已有 ADR，且包含替代方案和退出路径。

D. 架构决策记录（ADR）模板

D.1 ADR-XXX：决策标题

- 状态：提议 / 接受 / 废弃 / 被替代
- 日期：
- 决策人：

D.1.1 背景

需要解决的问题、约束和质量属性是什么？

D.1.2 备选方案

列出可行方案及其优缺点，不把当前偏好写成唯一选项。

D.1.3 决策

选择什么，为什么现在选择?

D.1.4 后果

带来哪些收益、成本、风险、技术债和组织影响?

D.1.5 验证与退出

用什么指标验证？什么条件下重新评估或退出？

E. Agent 权限矩阵模板

身份/角色	数据域	读取范围	字段/行限制	允许工具	允许动作	审批条件	审计要求

权限评审应分别覆盖用户、Agent 身份、服务账号和工具执行身份，并验证授权是否在检索、生成、缓存和行动阶段持续生效。

F. 高级岗位能力映射

书中产物	AI 数据产品经理	AI 数据架构师	首席/企业数据架构
场景优先级矩阵	价值判断、需求管理	可行性与依赖分析	投资组合与战略对齐
AI 数据产品 PRD	用户、指标、路线图	非功能需求、数据契约	跨域标准与治理原则
AI Ready 评估表	产品成熟度	能力差距与演进路径	企业级能力地图
参考架构	产品能力边界	组件、接口、数据流	目标架构与原则
ADR	产品取舍	技术权衡与后果	架构治理和例外管理
权限与审计矩阵	风险与体验平衡	安全控制落地	合规与企业政策
评估与反馈方案	产品效果与增长	端到端可观测性	价值治理与持续改进
90 天路线图	发布与运营	迁移和工程计划	组织协同与投资节奏

面试表达应围绕“背景—约束—选择—权衡—结果—复盘”，而不是只罗列技术名词。

参考文献