

“**Deep Dive: Creating Your First Data Connection**”(深入解析:创建你的第一个数据连接)是一门发布在 learn.palantir.com 上的实战进阶课程,由 **Ontologize** 团队(由前 Palantir 工程师组成)提供教学 1, 2。

该课程被视为用户在 Palantir Foundry 中获取价值的第一步,因为它教授如何将数据从外部系统引入平台 1。以下是根据来源对该课程标题及其核心内容的详细解释:

1. 核心工具:Data Connection 应用

Data Connection 是 Foundry 中专门用于将外部源系统的数据同步到平台内部,或者将清洗后的数据导回到源系统的核心工具 3。

- 功能定位:每当涉及数据在 Foundry 与外部系统(如云存储、数据库或 API)之间移动时,所有的配置都在此应用中完成 3, 4。
- 数据资产化:该工具最终将外部数据转化为 Foundry 中的原始数据集(**Raw Data Sets**),供后续的 Pipeline Builder 或分析工具使用 5。

2. 两个关键概念:源 (Sources) 与 同步 (Syncs)

课程强调了理解这两者之间关系的必要性,它们是构建连接的基础架构:

- 源 (**Sources**):存储如何连接到外部系统的指令 6, 7。它包含 URL、端口号以及安全凭据(如 API 密钥、机密令牌等)6, 7。源具备内置的机密管理器(**Secrets Manager**),确保敏感信息被加密存储 6。
- 同步 (**Syncs**):存储如何从特定源交换数据的指令 4, 6。例如,从一个源中提取哪个特定的表或文件夹 4。
- 关系:一个“源”可以拥有多个“同步”,但一个“同步”只能属于一个“源”,形成了一对多的关系 4。

3. 连接方式:直接连接 vs. 代理 (Agents)

课程详细解释了根据网络环境选择连接技术的标准:

- 直接连接 (**Direct Connection**):用于连接可通过互联网访问的数据源,如 REST API 或云存储 8, 9。其优势在于无需维护额外的软件,具备出色的运行时间和性能 9。
- 代理 (**Agents**):当数据源位于公司内部网络或防火墙后,无法通过公网访问时,需要部署代理软件 8, 10。代理通过单向加密连接与 Foundry 通信,确保企业内网的安全姿态 7, 8。

4. 数据保护:出口策略 (Egress Policies)

由于连接外部系统涉及信息安全,Foundry 默认锁定所有出站请求 11。

- 出口策略 (**Egress Policies**):这是允许 Foundry 向外部数据库或 API 发出请求的防火墙许可 4, 11。
- 权限要求:配置此类策略通常需要具备**信息安全官 (Information Security Officer)**角色的 elevated 权限 12, 13。

5. 三大核心实战场景

该课程之所以称为“Deep Dive”,是因为它涵盖了工业界最常见的三种数据源类型:

- 文件存储 (**AWS S3**):学习如何连接 S3 桶并摄取 .csv 或其他格式的文件,此流程同样适用于 Azure Blob 或 Google Cloud 存储 9, 14。

- 关系型数据库 (**PostgreSQL**): 演示如何通过 JDBC 协议连接数据库 15, 16。在底层, 这实际上是通过 **SQL** 查询来拉取数据的, 对于熟悉 SQL 的非数据工程师也非常友好 17。
- **REST API**: 展示如何通过**代码存储库 (Code Repositories)**以编程方式调用 API 接口, 并使用 Python 转换逻辑处理返回的 JSON 数据 16, 18, 19。

6. 数据溯源与管理

- 数据血缘 (**Data Lineage**): 课程展示了摄取后的数据集在 Data Lineage 应用中的样子 20, 21。用户可以清晰地看到数据来源的源、同步以及背后允许其通信的出口策略, 实现了端到端的数据追溯 20, 21。
- 自动化调度: 用户还可以为这些同步任务设置计划任务 (**Schedules**), 例如每天凌晨自动抓取最新数据 20, 22。

总结来说, 这个标题代表了一次从安全配置到数据落地的全流程实践。它不仅教导用户如何操作界面, 更深入解析了如何通过源管理、安全策略配置和代码集成, 在 Foundry 中建立起稳健、合规且自动化的企业级数据摄取链路 3, 22, 23。