“**Entity Extraction in Pipeline Builder**”（在 Pipeline Builder 中进行实体提取）是 **Ontologize** 团队提供的一门实战技术教程。该教程详细演示了如何在 Palantir Foundry 的 **Pipeline Builder** 中，利用大语言模型（LLM）从非结构化文档（如 PDF）中自动识别并提取特定的信息项（实体）1, 2。
以下是基于资料对该标题所涵盖内容的详细解释：

1. 核心定义与工具定位

- **Pipeline Builder**：Foundry 的主要数据集成和转换应用，提供低代码环境来构建数据流 1, 3。
- **Use LLM** 节点：Pipeline Builder 中的一个核心功能节点。它预设了六种不同的模板（分类、情感分析、摘要、实体提取、翻译、空提示词），帮助用户更轻松地编写提示词 2。
- 实体提取 **(Entity Extraction)**：当用户需要从大量文本中提取特定元素（如客户姓名、项目名称、日期或特定术语）时，应选择此模式 4。

2. 核心工作流程

该标题涉及的工作流展示了如何将 LLM 的推理能力无缝集成到数据管道中：

- 数据接入（**Media Set**）：教程首先将 PDF 文档上传并转化为 **Media Set**（媒体集）资源 3。值得注意的是，用户不需要先手动提取 PDF 文本，可以直接将媒体引用（Media Reference）传入 LLM 节点进行处理 2, 5。
- 配置提取逻辑：
- 上下文定义：设定提取的背景信息（例如：地热能源项目）5。
- 定义实体类型：明确要提取的目标。在教程的案例中，包括“地点（Place）”、“项目（Project）”和“引用论文（Cited paper）” 5, 6。
- 模型选择与提示词生成：用户可以选择不同的模型（如 **Gemini 2.0 Flash** 或 **GPT-4**）5, 6。系统会根据填写的模板自动生成系统提示词（指令）和任务提示词 6。

3. 技术进阶与数据结构化

为了使提取的结果具备业务可用性，教程强调了以下高级操作：

- 从字符串到数组（**Array of Strings**）：为了获取文档中出现的所有相关实体（而不仅仅是第一个），必须将输出类型从单一的 String 更改为 **Array** 6, 7。
- 结构化输出解析：LLM 返回的结果通常是一个复杂的结构体（**Struct**）7。教程教授如何使用 **“Extract many struct fields”**（提取多个结构体字段）转换操作，将提取出的“地点”、“项目”和“论文”分别转化为独立的数据列 8。
- 端到端自动化：最终生成的带实体信息的表可以部署为标准的数据集（Dataset），直接用于后续的分析或作为本体（Ontology）对象的支撑 9。

4. 业务应用价值

该技术通过“地热能源技术报告”的案例，展示了如何高效地导航复杂的专业文献库 1。其核心价值在于：

- 替代人工阅读：无需人工逐页查阅即可快速获取关键信息点 4。
- 提升数据质量：将杂乱的文本信息转化为结构化、可搜索的字段，便于后续的统计分析和知识图谱构建 6, 9。

总结来说，这个标题代表了在数据生产线中集成 **AI** 的前沿实践。它教导用户如何利用 Foundry 的内置 AI 功能，将非结构化的 PDF 文档规模化地转化为结构清晰、属性完备的行业数据库 1, 10。