“**Build with Us | Deep Dive: Building Your First Pipeline**”是发布在 learn.palantir.com 上的一门深度实战课程，由 **Ontologize** 团队制作 1, 2。该课程的核心目标是教授用户如何使用 **Pipeline Builder**——Palantir Foundry 的无代码、生产级数据管道构建工具，将原始数据转化为可用的清洗后数据资产 2-4。
以下是根据来源对该课程内容的详细解释：

1. 核心定位与工具选择

- 工具定位：Pipeline Builder 是 Foundry 中用于构建数据管道的低代码/无代码方案，而 Code Repositories 则对应高代码（Pro-code）方案 3, 4。
- 端到端流程：一个完整的数据管道通常涉及从外部系统接入、原始数据处理、清洗、连接，最终生成能够支撑本体（**Ontology**）、分析或模型的“黄金数据” 3, 4。

2. 数据准备与预处理 (Data Ingestion & Pre-processing)

课程模拟了真实的数据集成场景，通过上传不同格式的文件来启动流程：

- 多格式接入：用户需要下载并上传 **CSV**、**Parquet**（针对 Spark 优化的列式存储 ）和 **JSON** 格式的原始数据 5-7。
- 处理无模式数据：对于 JSON 等没有预设模式（Schema）的非结构化数据，课程演示了如何在不编写代码的情况下提取行信息、解析字典结构并生成表格式视图 8-10。

3. 数据清洗与高级转换 (Cleaning & Advanced Transforms)

课程详细涵盖了数据治理中的常见清洗动作：

- 基础转换：包括去除首尾空格（**Trim Whitespace**）、类型转换（**Cast**）（例如将字符串格式的价格转换为 Double 类型以便计算）11-14。
- 复杂结构处理：教授如何平铺结构体（**Flatten Struct**）。例如，将存储在 JSON 字典中的地址字段（街道、城市、州）提取出来，并重新**拼接字符串（Concatenate）**成人类可读的地址格式 15-21。
- **AI** 助手集成：演示了利用 Foundry 的 **AI 助手（AIP Assist）**通过自然语言描述来生成复杂的转换逻辑，例如自动识别并生成时间戳的格式字符串 22-25。

4. 数据集成与问题排查 (Data Integration & Debugging)

- 对象连接（**Joins**）：将交易数据（Transactions）与产品（Products）和客户（Customers）数据通过**左连接（Left Join）**进行关联 26-29。
- 识别“坏连接（**Bad Join**）”：这是课程的一个重点。当连接后的记录数异常增加（例如从 50 行激增至 172 行）时，引导用户使用 **Contour** 的直方图进行钻取分析，发现 ID 不唯一导致的重复问题，并学习通过更换连接键（如使用 Variation ID 替代 Product ID）来修复逻辑 30-37。

5. 业务逻辑派生与聚合 (Aggregations)

- 衍生指标计算：通过**乘法转换（Multiply）**计算每笔交易的收入（单价 × 数量），并命名为 Revenue 38, 39。
- 客户终身价值 **(CLV)**：使用聚合转换（**Aggregate/Group By**），按客户 ID 分组并计算收入总和，从而生成对业务具有重要运营价值的“客户终身价值”数据集 39-42。

6. 部署与管道维护 (Deployment & Maintainability)

- 实例化输出（**Materializing Outputs**）：强调 Pipeline Builder 中的节点只是中间结果，必须添加“新数据集”输出并执行**部署（Deploy）**动作，数据才会在平台中实际构建并可用 34, 43-46。
- 最佳实践与进阶概念：
- 管道分层（**Segmentation**）：建议按照原始（Raw）、清洗（Clean）、丰富（Enriched）和本体（Ontology）等阶段对管道进行拆分，以提高可读性和可维护性 47-50。
- 数据预期（**Data Expectations**）：设置主动监控检查（如检查主键唯一性、行数范围），一旦不符合规则可触发警告或中断构建，防止坏数据传播 51。
- 增量转换（**Incremental Transforms**）：对于大规模数据，仅处理新流入的行以减少计算负载并提高速度 51, 52。

通过本课程，用户能够建立起**“将杂乱的原始文件转变为工业级、受治理的数据资产”**的全局观，并熟练掌握 Pipeline Builder 的图形化操作界面 1, 53。