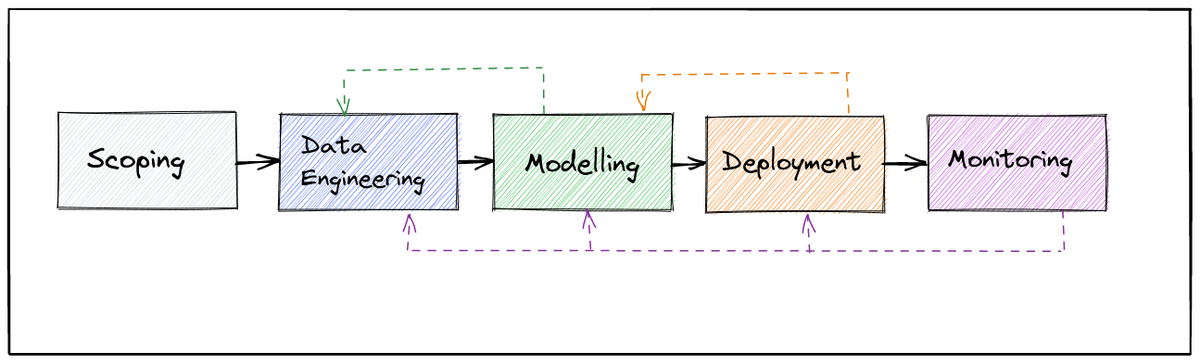
**MLOps**

1. 什么是MLOps？

机器学习操作涉及一套流程，或者说是为将ML模型部署到生产环境而实施的一系列步骤。（之后我们要讲这个全过程）这些步骤确保模型具备泛化能力，并准确执行。

创建一个能够基于输入的数据，训练得到模型是很容易的。然而，创建一个可靠、快速、准确并能被大量用户使用的ML模型是很困难的。（即模型的工程化是难的）

2. ML项目生命周期



范围界定：我们定义项目，检查问题是否需要机器学习来解决。执行需求工程，检查相关数据是否可用。核实数据是否无偏见，是否反映了真实世界的使用情况。

数据工程：这一阶段包括收集数据、建立基线、清理数据、格式化数据、标记和组织数据。

建模：现在我们来到了编码部分，在这里我们创建了ML模型。我们用处理过的数据来训练模型。进行误差分析，定义误差测量，并跟踪模型性能。

部署：在这里我们对模型进行打包，根据需要在云端或边缘设备上进行部署。包装可以是--用暴露REST或gRPC端点的API服务器包装的模型，部署在云基础设施上的docker容器，部署在无服务器的云平台上，或基于边缘模型的移动应用程序。

监测：一旦部署完成，我们就依靠监测基础设施来帮助我们维护和更新模型。这个阶段有以下几个部分。

* 监测我们部署的基础设施，包括负载、使用、存储和健康。这告诉我们关于ML模型部署的环境。
* 监测模型的性能、准确性、损失、偏差和数据漂移。这告诉我们，该模型的性能是否符合预期，对现实世界的情况是否有效。

3. ML基础设施

数据收集：这一步涉及从各种来源收集数据。ML模型需要大量的数据来学习。数据收集主要完成整合与问题相关的各种原始数据。

数据验证：再这一步，我们检查数据的有效性，如果收集的数据是最新的、可靠的，并反映了真实的世界，它是否已是恰当的的可被利用的格式，数据的结构是否正确需要进行验证。

特征提取：为模型选择最佳的特征用于预测。换句话说，模型可能不需要全部的数据来发现模式，一些列或部分数据可能根本就不需要被使用。有些模型在删除/丢弃一些列时会表现更好。因此，通常会按重要性对特征进行排序，重要性高的特征被包括在内，较低的或接近零的特征被放弃。

配置：这一步涉及设置通信协议、系统集成以及管道中的各种组件应该如何相互对话。希望数据管道连接到数据库、希望ML模型以适当的访问方式连接到数据库、希望模型以某种方式暴露预测端点、希望模型输入以某种方式格式化，系统所需的所有必要的配置都需要恰当地确定并记录。

ML代码：在这个阶段完成开发一个基础模型，它可以从数据中学习和预测。有大量的ML库，支持多种语言，例如：Pytorch、sklearn、keras、fast-ai等。一旦有了一个模型，就可以开始通过调整超参数，测试不同的学习方法来提高它的性能，直到我们确信这个模型的性能比它以前的版本要好。

机器资源管理：这一步涉及ML模型的资源规划。通常，ML模型在CPU、内存和存储方面需要大量资源。深度学习模型依赖GPU和TPU进行计算。训练ML模型涉及时间和金钱方面的成本。较慢的CPU涉及更多的时间，强大的CPU则价格较高。模型越大，要投资的存储就越大。

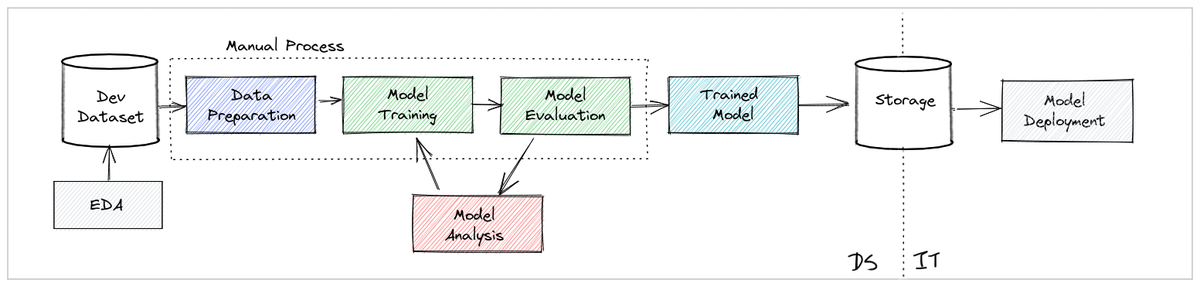
分析工具：一旦模型准备好了，模型的性能是否达到了标准，主要在该阶段通过模型分析确定。如何计算损失、应该使用什么误差测量、如何检查模型是否漂移（泛化能力）、预测结果是否正确、模型是否发生过度拟合或欠拟合，这些问题都可以通过实现相应的模型的分析包和误差测量完成。

项目管理工具：实现跟踪一个ML项目。在处理庞大的数据、功能、ML代码、资源管理时，很容易迷失方向，可以使用互联网上提供的已有项目管理工具进行辅助管理。

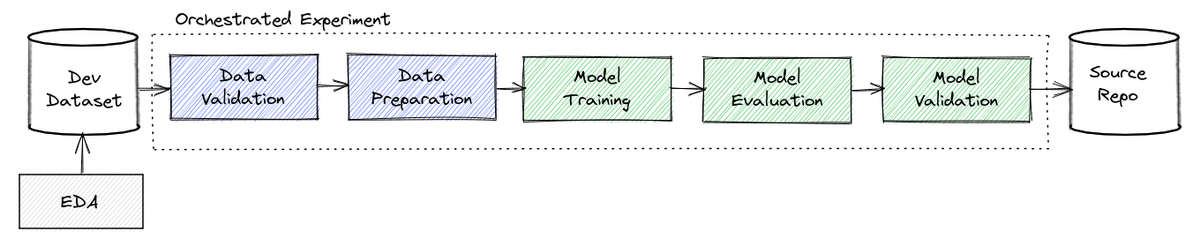
服务基础设施：一旦模型被开发、测试并准备就绪，就需要将其部署到用户可以访问的地方，大部分的模型都部署在云上。如AWS、GCP、Azure等公共云。

监控：通过实现一个监控系统来观察部署的模型和它所运行的系统。收集模型日志、用户访问日志和预测日志，这些日志将有助于维护模型。可用的监控解决方案如greylog、elasticstack、fluentd等。云供应商通常会提供开箱即用的监控系统。

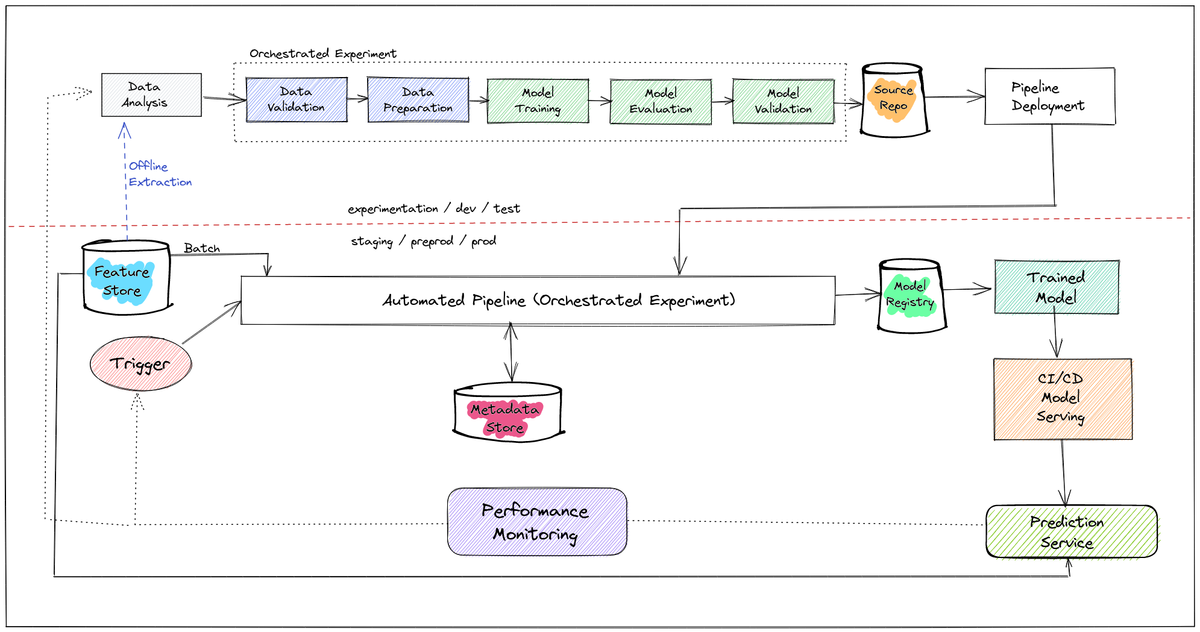
Level-0 MLOps过程：手动实验过程



Leve-1 MLOps过程：创建了一个自动化管道，以验证数据，进行准备，并训练模型



Leve-2 MLOps过程：



在红线以上所做的是Level-1 MLOps过程，整个Orchestrated Experiment是自动的ML管道的一部分。引入一个特征库，它从各种来源获取数据，并将数据转换为模型所需的特征，ML管线分批使用存储的数据。ML管道连接到一个元数据存储，可以把它看作是记账，因为这样就可以不需要手动训练模型，元数据存储中包含管道中每个阶段的记录，一旦一个阶段完成，下一个阶段就会查找记录列表，找到前一个阶段的记录，并从该阶段继续。然后，这些模型被存储在一个模型注册中心。我们有一堆具有不同精度的模型存储在这里。根据需求，适当的模型被发送到CI/CD管道，该管道将其作为预测服务进行部署。经授权的用户能够在需要时访问预测服务。这个系统的性能受到监控。假设你有一堆新的转基因南瓜。你的模型并不清楚这一点。这是一个新的数据集，很有可能被错误地分类。这种性能的下降将设置一个触发器，这将导致在新数据上重新训练模型。这个循环还在继续。

4. MLOps工具

MLFlow、Seldon Core、Metaflow和Kubeflow Pipelines