数据管理检查清单

欢迎进入人工智能 (AI) 时代，这是一个需要机器学习和深度学习等数据密集型技术来经营业务的时代。要用好这些新的 AI 工具，您必须确保企业的数据“仓库”已准备就绪。

下面是一份检查列表，帮助您开始建立一个清洁的数据仓库，它包含数据管理的两个重要阶段 — 训练和推断。

做到以下几步，助您成为一名 AI 高手。要进一步了解如何从概念验证走向全面的 AI 生产和规模化，请参阅此 IDC 报告。*使用 AI 优化的基础架构加速并实施 AI 部署*。

训练

在为 AI 准备的训练阶段，您会开发出用于理解数据集的算法。您的主要关注点在于收集现有数据并利用 AI 学习新功能。

- [ ] 找出您希望通过 AI 来解决的特定业务问题（从较小的项目开始以便您从中积累经验）
- [ ] 从相关资源中找到可以解决该问题的数据（它极有可能不在一处）
- [ ] 用元数据标签准备数据，以大幅缩短查找相关数据所需的时间
- [ ] 确保在您使用的所有数据集中正确同步和关联数据（包括同步时间）
- [ ] 标记客户敏感的数据及其他私密数据，确保数据安全无虞并遵守所有相关规定和法规（元数据标记流程对此有帮助）
- [ ] 为您使用的数据类型选择正确的开发环境和格式（即图像、视频、自由文本和音频通常都有自己独特的环境）
- [ ] 从资源库中提取数据集，并将数据集带入开发环境
- [ ] 将数据分成两组，以改进模型开发流程（一组放在名为“训练”的文件夹内，另一组放在名为“测试”的文件夹内）
- [ ] 通过跟踪数据的来源来维护数据的可追溯性（考虑使用可自动执行该流程的工具）
- [ ] 执行基本的数据清理任务，为构建模型准备数据（例如补充缺失的数据条目，删除空条目）
- [ ] 使用您已知道预测活动答案的数据子集样本（这就是所谓的“训练集”），并确定准备数据以进行预测所需的预处理步骤
- [ ] 运用此训练集的知识计算准确性得分，让您有信心将同样的模型用于其模型从未明确训练的新数据

推断

开发出针对业务问题对症下药的模型之后，您将从训练阶段进入推断阶段。在此阶段内，您可采用成功的模型并将它用于新数据，这个过程同样需要持续的数据管理。

- [ ] 找到邻近您数据的 AI 模型，以缩短延迟，降低带宽要求并提升整体模型绩效
- [ ] 开发高效的数据管道流程，将元数据标签应用到进入的数据中，以收集新数据并将它们用于改进模型
- [ ] 以关联和同步的方式标记数据（例如，如果数据按时间排列顺序，则可以在数据集进行同步或者挑选一个字段（例如客户名称）以关联陆续进入的所有数据）
- [ ] 制定一份长期的数据生命周期存储计划，用于管理新增数据和归档数据的量和速度
- [ ] 考虑聘请一位首席数据官来维护企业的数据管理，以实施未来的 AI、深度学习及其他数据驱动的项目

IBM