数据整理清单

欢迎来到人工智能 (AI) 时代，您将依靠数据密集型技术（如机器学习和深度学习）来开展业务运营。要利用这些新型 AI 工具，您需要确保组织的数据仓库井然有序。

以下清单可帮助您开始清理数据仓库，数据整理分为两个主要阶段 - 训练和推理。

遵循这些步骤可帮助您的成为一名 AI 大师。要深入了解在 AI 旅程中如何从概念证明阶段迈向全面投产和规模应用阶段，请查看这份 IDC 报告：*利用 AI 优化的基础架构加速实现 AI 部署*。

训练

在为 AI 做准备的训练阶段，您将开发各种算法来理解数据集。您的主要问题是收集现有数据并利用 AI 来学习新功能。

- [ ] 弄清您想要利用 AI 解决的具体业务问题（从较小的项目开始将有利于学习）
- [ ] 从相关来源中找出能解决该问题的数据（这些数据极有可能不在单一位置）
- [ ] 使用元数据标签来准备数据，显著缩短查找相关数据所需的时间
- [ ] 确保数据正确同步并在您将使用的所有数据集之间互相链接（包括时间同步）
- [ ] 标记任何客户敏感型数据和其他私有数据，确保数据绝对安全并遵守所有相应的监管和法规条例（元数据标签过程有助于实现此目的）
- [ ] 针对您使用的数据类型以及数据格式选择合适的开发环境（例如图像、视频、自由格式文本和音频，通常每一类都有一个对应的环境）
- [ ] 将数据集从存储库中提取出来并传送到您的开发环境。
- [ ] 将数据分为两组，来帮助改进您的模型开发过程（将一组放置在名为“train”的文件夹中，另一组保存在名为“test”的文件夹中）
- [ ] 持续跟踪数据的来源，保持数据可跟踪性（考虑使用工具来帮助自动执行这一流程）
- [ ] 执行基本数据清理任务，以准备好数据来构建模型（例如，包括填充缺失的数据条目并除去空条目）
- [ ] 使用您已知其预测活动答案的数据子集样本（称为“训练集”），并确定准备数据进行预测所需的所有预处理步骤
- [ ] 使用此训练集的知识来计算准确得分，这会让您有信心将相同的模型应用于模型从未明确训练过的新数据

推理

开发出模型来解决您的业务问题后，您将从训练阶段进入推理阶段。在此阶段，您会将这一成功模型应用于新数据，期间需要持续开展数据整理工作。

- [ ] 找到与您的数据最接近的 AI 模型，以缩短延迟、降低带宽需求并改进模型的整体性能
- [ ] 开发高效的数据管道流程，并在数据传入时添加元数据标签，这样即可收集新数据并将其用于持续增强模型
- [ ] 以互相链接且同步的方式来标记数据（例如，如果数据按时间排序，可同步数据集或通过挑选一个字段（如客户名称）让传入的所有数据相互链接）
- [ ] 制定长期的数据生命周期存储计划，帮助在传入和归档数据时管理数据的数量和速度
- [ ] 考虑聘用一名首席数据官来持续管理组织的数据，以用于未来 AI、深度学习和其他数据驱动型项目

© Copyright IBM Corporation 2019

IBM