データ・ハウス・キーピング・チェックリスト

人工知能 (AI) の時代にようこそ。この時代では、機械学習やディープ・ラーニングといったデータ集約型テクノロジーを活用してビジネスを行います。こうした新しい AI ツールを利用するには、組織のデータ「ハウス」がきちんと整備されていることを確認する必要があります。

以下に、クリーンなデータ・ハウスを構築するために最初に確認すべきチェックリストを、トレーニングと推論というハウスキーピングの 2 つの主要な段階に分類して示します。

ここで挙げられた手順に従って、AI の達人を目指しましょう。AI を概念実証から完全な実稼働とスケールに移行するための洞察の詳細は、以下の IDC レポートをご覧ください。*Accelerate and Operationalize AI Deployments Using AI-Optimized Infrastructure* (*AI に最適化されたインフラストラクチャーを使用して AI の展開を加速し、運用する*)。

トレーニング

AI のためのトレーニング準備段階では、データ・セットを理解するためのアルゴリズムを開発します。ここでの主な懸念事項は、既存データの収集と、AI を使用した新機能の学習です。

- ☐ AI を使用して解決したいビジネスの特定の課題を見極めます (最初は学習のために小規模プロジェクトから始めます)。
- ☐ その課題を解決できるデータを適切なソースから見つけます (多くの場合、すべてのデータは 1 か所にはありません)。
- ☐ 適切なデータを見つけるための所要時間を大幅に短縮するために、メタデータ・タグを使用してデータを準備します。
- ☐ 使用するすべてのデータ・セットについて、データが適切に同期化およびリンクされていることを確認します (時間の同期化を含む)。
- ☐ 機密データのセキュリティーを確実に保護し、すべての適切なガバナンスと規制に準拠するように、顧客の秘密データや他の個人情報にフラグを付けます (メタデータのタグ付けプロセスが役立ちます)。
- ☐ 使用するデータの種類とその形式に適切な開発環境を選択します (つまり、イメージ、ビデオ、フリーフォーム・テキスト、オーディオは通常、それぞれ特定の環境を持ちます)。
- ☐ リポジトリーからデータ・セットを抽出し、開発環境に取り込みます。
- ☐ モデル開発プロセスを改善するために、データを 2 つのグループに分類します (一方のセットは「トレーニング」フォルダーに、他方のセットは「テスト」フォルダーに入れます)。
- ☐ どこの/どのソースからデータが来たかを追跡することで、データのトレーサビリティーを維持します (このプロセスを自動化するためのツールの使用を検討してください)。
- ☐ 基本的なデータ・ハイジーン・タスクを実行して、モデルの構築用にデータを準備します (例: 欠けているデータ項目を埋める、null 項目を削除する)。
- ☐ 予測活動の答えをすでに知っているデータのサブセット・サンプルを使用して (これは「トレーニング・セット」と呼ばれます)、予測を行うためのデータの準備に必要なすべての事前処理ステップを特定します。
- ☐ このトレーニング・セットの知識を使用して、精度スコアを計算します。これにより、同じモデルを、そのモデルが明示的にトレーニングされていない新しいデータに適用する自信が得られます。

推論

ビジネスの問題を解決するために有効なモデルを開発したら、トレーニングから推論に移行します。この段階では、開発に成功したモデルを新しいデータに適用します。このデータにも、継続的なハウスキーピングがいくつか必要となります。

- ☐ レイテンシーを短縮し、必要な帯域幅を減らし、モデルの全体的なパフォーマンスを改善するために、データの近くに AI モデルを配置します。
- ☐ 効率的なデータ・パイプライン・プロセスを開発して、蓄積されるデータにメタデータ・ラベリングを適用し、モデルを発展、進化させるために新しいデータを収集して使用できるようにします。
- ☐ リンクおよび同期化しながらデータをタグ付けします (例えば、データが時系列になっている場合は、蓄積される全データ上の 1 つのフィールド (顧客名など) を選択することで、データ・セットまたはリンクについて同期化できます。
- ☐ 蓄積およびアーカイブするデータのボリュームとスピードを管理する方法について、長期的なデータ・ライフサイクル・ストレージ計画を立てます。
- ☐ 将来の AI、ディープ・ラーニング、および他のデータ駆動型プロジェクトに向けて組織のデータ管理を維持するために、最高データ責任者を採用することを検討します。

IBM