Data Housekeeping Checkliste

Willkommen im Zeitalter künstlicher Intelligenz (KI oder AI), in dem datenintensive Technologien wie Machine Learning und Deep Learning unerlässliche Begleiter der Unternehmen geworden sind. Voraussetzung für eine erfolgreiche Nutzung der neuen KI-Tools ist eine gewisse „Ordnung im Datenhaus“.

Folgende Checkliste soll helfen, Ihr „Datenhaus“ aufzuräumen. Dabei werden zwei Hauptschritte des „Housekeepings“ unterschieden: Training und Inferenz.

Diese Schritte helfen Ihnen, erfolgreich in KI zu werden. In diesem IDC Bericht erhalten Sie weitere Einblicke dazu, wie Sie KI von der Proof-of-Concept-Phase im großen Stil produktiv nutzen können. *Accelerate and Operationalize AI Deployments Using AI-Optimized Infrastructure („Beschleunigung und operativer Einsatz von KI-Implementierungen basierend auf KI-optimierten Infrastrukturen“)*.

Schulung

Im Rahmen des Trainings zur Vorbereitung auf KI, werden Algorithmen entwickelt, um Datensätze verständlich zu machen. Die Hauptaufgabe besteht darin, vorhandene Daten zusammenzutragen und KI zum Erlernen neuer Fähigkeiten einzusetzen.

- [ ] Überlegen Sie sich ein bestimmtes Geschäftsproblem, das Sie mit KI lösen möchten (beginnen Sie mit kleineren Projekten, um schrittweise zu lernen)
- [ ] Wählen Sie die Daten aus den entsprechenden Quellen, die das Problem lösen können (höchstwahrscheinlich befinden sie sich nicht an einer Stelle)
- [ ] Bereiten Sie Ihre Daten mit Metadata Tags vor,um den Aufwand für die Auswahl der relevanten Daten erheblich zu reduzieren
- [ ] Stellen Sie sicher, dass die Daten in allen genutzten Datensätzen entsprechend synchronisiert und verknüpft sind (einschließlich Zeitsynchronisierung)
- [ ] Kennzeichnen Sie alle kundensensiblen und anderen geschützten Daten, um deren volle Sicherheit und die Einhaltung aller entsprechenden unternehmensspezifischen und rechtlichen Vorschriften zu gewährleisten (hierbei hilft der Metadata-Tagging-Prozess)
- [ ] Wählen Sie die geeignete Entwicklungsumgebung für die Art der Daten, die Sie nutzen und deren Format (d. h. Bild, Video, Frei-text und Audio haben normalerweise eine eigene Umgebung)
- [ ] Entnehmen Sie Datensätze aus Ihrem Archiv und übertragen Sie diese in Ihre Entwicklungsumgebung
- [ ] Teilen Sie die Daten in zwei Gruppen auf, um die Verbesserung des Modellentwicklungsprozesses zu unterstützen (eine Gruppe wird in einem Ordner „Training“ und eine andere in einem Ordner „Test“ aufbewahrt)
- [ ] Achten Sie auf die Rückverfolgbarkeit der Daten, indem Sie den Ursprung/ die Quelle Ihrer Daten überwachen (nutzen Sie ggf. dazu Tools zur Prozessautomatisierung)
- [ ] Führen Sie grundlegende Datenhygieneaufgaben aus, um die Daten für die Entwicklung eines Modells vorzubereiten (z. B. einschließlich Ergänzung fehlender Dateneinträge und Entfernen leerer Einträge)
- [ ] Nutzen Sie einen Auszug der Daten, von denen Sie bereits das Ergebnis der Vorhersageaktivität kennen (der so genannte „Trainingssatz“) und ermitteln Sie alle vorhergehende Vorbereitungsschritte, die notwendig sind, um die Daten für eine Vorhersage vorzubereiten
- [ ] Nutzen Sie Ihr Wissen aus diesem Trainingssatz, um Genauigkeitswerte zu berechnen, die Ihnen die Sicherheit geben, das gleiche Modelle auf neue Daten anzuwenden, für die das Modell nicht explizit trainiert wurde

Inferenz

Sobald ein Modell entwickelt wurde, das funktioniert, um ein Geschäftsproblem zu lösen, erfolgt der Schritt vom Training zur Inferenz. In dieser Phase wird das erfolgreiche Modell auf die neuen Daten angewendet. Hierbei ist ebenfalls fortlaufendes Data Housekeeping erforderlich.

- [ ] Um Latenzzeiten zu verkürzen, Bandbreitenanforderungen zu minimieren und die Gesamtleistung des Modells zu optimieren, sollten AI-Modell und Daten so nah wie möglich sein.
- [ ] Entwickeln Sie einen effizienten Daten-Pipeline-Prozess und nutzen Sie Metadata Labeling für eingehende Daten, sodassneue Daten erfasst und genutzt werden können, um das Modell immer weiter zu verbessern
- [ ] Kennzeichnen Sie Daten auf verknüpfte und synchronisierte Weise (z. B. kann bei zeitlich sortierten Daten eine Synchronisierung der Datensätze oder eine Verknüpfung durch Auswahl eines Feldes erfolgen — z. B. Kundenname — für alle eingehenden Daten)
- [ ] Entwickeln Sie einen langfristigen Aufbewahrungsplan, der den Lebensszyklus der Daten reflektiert und der Menge und Datengeschwindigkeit der eingehenden und zu archivierenden Daten berücksichtigt.
- [ ] Setzen Sie einen Chief Data Officer (Datenverantwortlichen) für das Datenmanagement im Unternehmen ein - für künftige KI-, Deep Learning- und andere datenbasierte Projekte.

IBM