MLOps als skalierbare System-Architektur

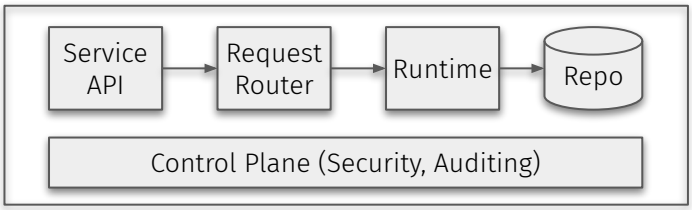
Einfache Prozess-Integration von AI Lösungen auf Basis einer durchdachten Plattform

Erfolgsfaktor Prozess-Integration

Eine **AI Lösung erzeugt nur Mehrwert,** wenn sie in die betrieblichen Prozessen integriert ist. Studien zeigen, dass Unternehmen, die AI in ihrer Strategie verankert haben, viel erfolgreicher sind, als jene, die AI lediglich zur Effizienzsteigerung einsetzen. AI Lösungen müssen deshalb so entwickelt sein, dass die **Integration in betriebliche (IT-)Prozesse schnell und technisch unkompliziert** möglich ist. Diese Anforderung wird am besten erfüllt durch ein standardisiertes Vorgehen, gestützt auf eine dafür ausgelegte MLOps Plattform.

MLOps als System-Architektur

Die omega-ml MLOps Plattform besteht aus folgenden Elementen. Diese Architektur ist mit Standard-Technologie umsetzbar, flexibel und kosteneffizient. Dank einem hohen Grad an Standardisierung gestaltet sich das Deployment einfach und ohne starre Abhängigkeiten. Die Plattform ist für unterschiedliche Szenarien horizontal und vertikal skalierbar.



Diese bewährte Architektur umfasst alle Funktionen:

- ➜ **Service API** ist die zentrale Schnittstelle für Dritt-Anwendungen, z.B. via REST API
- ➜ **Request Router** sichert jeden Request ab und weist ihn der geeigneten Runtime zur Ausführung zu
- ➜ **Runtime** verfügt über die erforderlichen Ressourcen je nach Modelltyp (z.B. Python-Libraries, Memory und CPU für klassisches ML, ggf. GPU für Generative AI) und ist horizontal und automatisch skalierbar
- ➜ **Repository** stellt Modelle bereit und ermöglicht den geschützten Zugriff auf Daten aus Drittquellen (z.B. Data Lake, DWH, operative Systeme)
- ➜ **Control Plane** übernimmt die Verwaltung aller Ressourcen, den Zugriffsschutz, das Monitoring und stellt die Nachvollziehbarkeit aller Aktivitäten sicher.

omega-ml verwendet Flask bzw. Django für das Service-API, RabbitMQ als Request-Router, Python Celery für die Runtime und MongoDB für das Repository. Die Control-Plane wird in Standard-IT Komponenten wie z.B. Kubernetes, Keycloak integriert.

Weshalb MLOps als Plattform?

Die Komplexität von AI Lösungen wird oft unterschätzt, weil die Anforderung auf den ersten Blick unspektakulär erscheint: Es gilt, den an sich einfachen IT Service - “predict” - zu paketieren und als REST-API Schnittstelle in einen betrieblichen Prozess zu integrieren. Eine MLOps Plattform bietet jedoch viele Vorteile.

Aspekt	Einzelner Service	MLOps Plattform
Fähigkeiten	Data Science, DevOps, Software-Engineering für verteilte Systeme	Data Science
Technologie-Stack	Web- und REST API wie Flask, FastAPI oder Django, Docker, Kubernetes, etc.	Standard Python
Skalierbarkeit	Tief, nicht einfach zu erreichen	Hoch, automatisiert
Sicherheit	Komplex, für jeden Service neu	Standardisiert, für alle Services gleich
Governance	Aufwändig, hoher Abstimmungsbedarf	Einheitlich, einmalige Abstimmung

Umsetzung in 5 Schritten

Unsere praxiserprobten Schritte zur erfolgreichen Einführung von MLOps eigenen sich für alle typischen Data Science Teams mit 1 bis 10 Personen:

(1) Make oder Buy? **omega-ml** ✅
Eine eigene Entwicklung bietet hohe Flexibilität, ist jedoch komplex und angesichts der Verfügbarkeit von standardisierten Plattformen vergleichsweise nicht kosteneffizienz.

(2) Wahl einer integrierten MLOps Plattform ✅
Viele MLOps Angebote bieten keine integrierte Plattform, sondern nur Teilkomponenten.

(3) Proof of Concept ✅
Ein technischer Test anhand eines realistischen Szenarios zeigt schnell auf, wo Nutzen und Herausforderungen liegen.

(4) Operationalisierung ✅
Mit Integration in die bestehende IT-Architektur, DevOps-Prozesse und Ihre Sicherheits-Umgebung steht die MLOps Plattform für den produktiven Betrieb zur Verfügung.

(5) Schulung und Skalierung nach Bedarf ✅
Um die Plattform effizient zu nutzen, ist die Einführung und Schulung von Data Scientists und DevOps Spezialisten ein effizientes Mittel.