



50% remote – Cloud Platform Engineer (Automation) (m/w/d)

Für unseren Kunden suchen wir einen Cloud Platform Engineer (m/w/d).

Rahmendaten:

Start: asap

Laufzeit: bis 31.12.2027 (Option auf Verlängerung)

Auslastung: VZ

Einsatzort: Remote & Köln (nach Absprache remote-Anteile von 50% möglich)

Aufgabenbeschreibung

In dieser Rolle verbindest du AI Engineering, Data Engineering & Cloud Architecture. Gemeinsam mit dem Team entwickelst du auf Data Azure Lösungen, die Daten, Large Language Models und interne Systeme miteinander verbinden.

Aufgaben:

- Konzeption & Entwicklung von produktiven Agentic-AI-Lösungen auf Databricks und Azure – vom ersten Prototyp bis zum skalierbaren Einsatz
- Entwicklung von KI-Agenten und Multi-Agent-Workflows, die Large Language Models, Unternehmensdaten, APIs und interne Systeme miteinander verbinden.
- Umsetzung von RAG-, Tool-Calling- und Workflow-Architekturen und Entwicklung von wiederverwendbaren Komponenten für Prompts, Agenten und Datenzugriffe.
- Integration von Azure OpenAI & weitere geeignete Foundation Models in die Databricks-Plattform sowie in bestehende Applikationen und Datenlandschaften.
- Entwicklung von Daten- und Verarbeitungspipelines auf Databricks, unter anderem mit Python, SQL, Spark und Delta Lake, und schaffen eine verlässliche Datenbasis für KI-Anwendungen.
- Nutzung von Databricks Mosaic AI und Mlflow, um Modelle, Prompts und Agenten zu entwickeln, zu versionieren, zu evaluieren, bereitzustellen und zu überwachen.
- Automatisierung von Daten- und KI-gestützten Prozessen mit Databricks Workflows und integrierst diese in bestehende Unternehmens- und Systemlandschaften.
- Gestaltung von erforderlichen Azure- und Databricks-Architekturen für einen sicheren, skalierbaren und kosteneffizienten Betrieb.
- Etablierung von CI/CD- und MLOps-Prozessen für Datenpipelines, Modelle und Agenten. Dazu gehören automatisierte Tests, Deployment Pipelines, Infrastructure as Code und die Versionierung relevanter Artefakte.
- Entwicklung von Qualitäts- und Evaluationsverfahren für KI-Anwendungen, beispielsweise für Antwortqualität, Groundedness, Sicherheit, Kosten und Zuverlässigkeit.
- Sicherstellung von Betrieb durch Unity Catalog, Secrets Management, Managed Identities, RBAC, Private Endpoints, Logging und Monitoring.
- Anbindung von internen Datenquellen, Systemen und APIs und Unterstützung der Fachbereiche bei der technischen Umsetzung konkreter Automatisierungs-UseCases.
- Zusammenarbeit mit Product Owner, Fachbereichen, Data- und Cloud-Teams sowie Security, Privacy und Governance
- Dokumentation von Architekturen, Datenflüssen, Agentenlogik, Betriebsprozessen und Security Controls so, dass ein nachhaltiger Übergang zum produktiven Betrieb möglich ist.

Profilanforderungen

- Mehrjährige Erfahrung als AI Engineer, Machine Learning Engineer, Data Engineer, Cloud Engineer oder in einer vergleichbaren Rolle.

- Praktische Erfahrung bei der Entwicklung von Anwendungen mit Large Language Models, idealerweise mit RAG, Embeddings, Vector Search oder agentenbasierten Architekturen.
- Sehr gute Kenntnisse in Python und SQL sowie Erfahrung mit Software-EngineeringPraktiken wie Git, automatisierten Tests, Code Reviews und modularer Softwareentwicklung.
- Fundierte Erfahrung mit Databricks , beispielsweise mit Spark, Delta Lake, Workflows, Unity Catalog, MLflow oder Mosaic AI.
- Gute Kenntnisse der Microsoft-Azure-Plattform , insbesondere Azure OpenAI, Managed Identities, Key Vault, Storage, Networking und
- Erfahrung mit der produktiven Integration von APIs, Datenquellen & Unternehmenssystemen sowie ein gutes Verständnis verteilter & ereignisgetriebener Architekturen.
- Praktische Erfahrung mit CI/CD, MLOps und Infrastructure as Code , beispielsweise mit GitHub Actions oder Azure DevOps sowie Terraform oder Bicep.
- Solides Verständnis von Cloud Security, Datenschutz und AI Governance , insbesondere von Berechtigungskonzepten, Datenklassifizierung, sicheren Modellzugriffen & dem Schutz sensibler Informationen.
- Erfahrung mit dem Monitoring produktiver KI-Lösungen und der systematischen Bewertung von Prompt-, Retrieval- und Agentenqualitätsmetriken.
- Fähigkeit, Anforderungen aus den Fachbereichen in robuste, wartbare und skalierbare technische Lösungen zu übersetzen.
- Freude an interdisziplinärer Zusammenarbeit, eigenverantwortlichem Arbeiten und der verständlichen Vermittlung komplexer technischer Zusammenhänge.

Nice to Have

- Erfahrung mit gängigen Agenten- und LLM-Frameworks sowie mit strukturierten Ausgaben, Guardrails und Human-in-the-Loop-Prozessen.
 - Kenntnisse in Docker und Kubernetes & im Betrieb containerisierter KIAnwendungen und APIs.
 - Erfahrung mit Event Streaming, beispielsweise auf Basis von Azure Event Hubs oder Apache Kafka .
 - Verständnis für die Kosten-, Latenz- & Skalierungsoptimierung LLM-basierter Anwendungen.
 - Erfahrung in regulierten oder datenschutzsensiblen Unternehmensumfeldern.
-