



Von Cloud- Kostentransparenz zu strategischer Steuerung

Christina Thumser | Head of Infrastructure Services | T-Systems
RS GmbH

T Systems

Let's power
higher performance

Introduction

About Christina:

Christina Thumser

Head of Infrastructure Services



About T-Systems Road User Services GmbH:

- Subsidiary of T-Systems International GmbH
- Development, implementation, and operation of intelligent tolling solutions



T Systems Let's power
higher performance

Reifegradmodell Cloud-Kostenmanagement

01 Transparenz

Wir sehen die Kosten!

- Toolgestützte Transparenz (z. B. mit IBM Cloudability)
- Kosten sichtbar & zuordenbar
- Reporting & Dashboards

02 Steuerung

Wir entscheiden bewusst!

- Cloud-Kosten als Investitionen
- Klare Produktverantwortung
- Priorisierung statt pauschaler Optimierung

03 Business Value

Wir steuern Wertbeiträge!

- Verknüpfung von Kosten & Business-Zielen
- Governance & FinOps als Betriebsmodell
- Messbarer Beitrag zum Geschäftserfolg

01 Transparenz

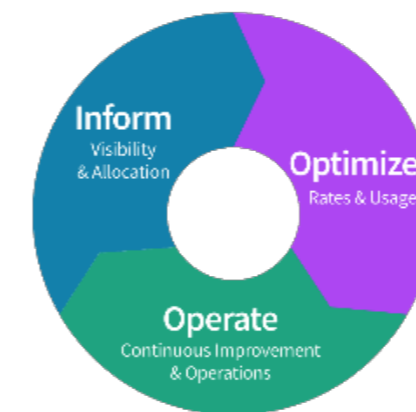
Die Herausforderungen:

- Wir können nur steuern, was wir auch sehen können.
- Skalierbarkeit bringt Flexibilität, aber auch dynamische Kosten.



Der Weg zur Transparenz:

- Die Wahl der richtigen Lösung

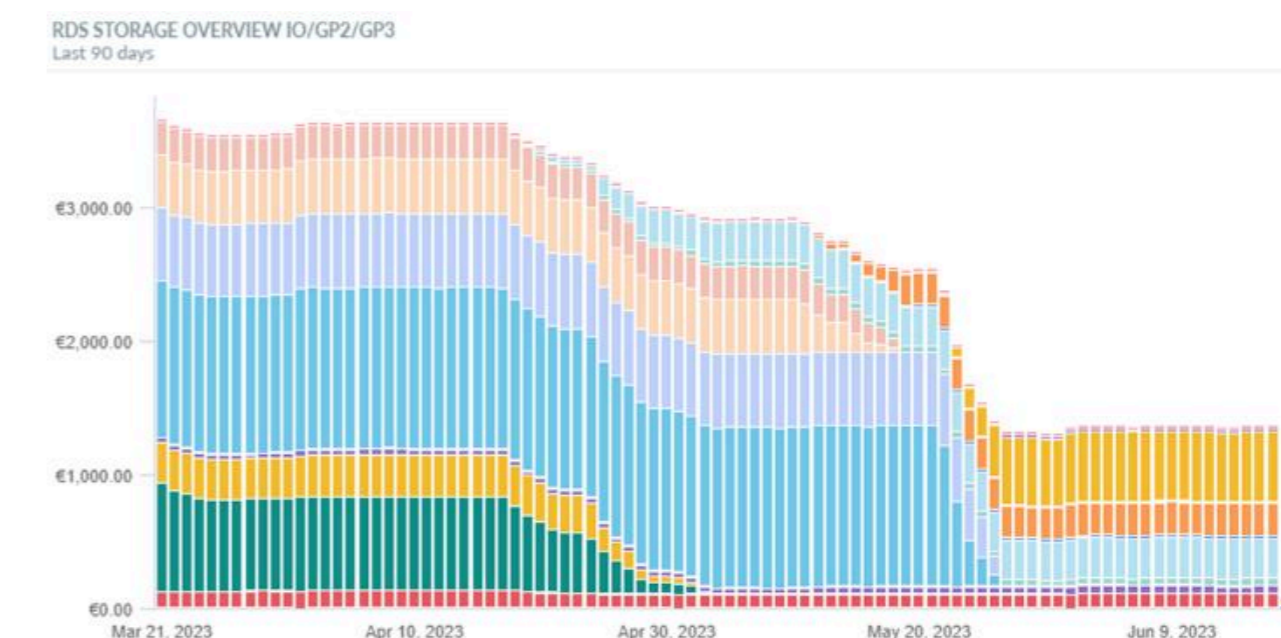


IBM Cloudability

- ✔ Real-time, dynamic cost visibility
- ✔ Ability to allocate costs by project, service, and team
- ✔ Support for FinOps best practices
- ✔ Integration with AWS
- ✔ Proven experience and maturity in cloud cost management

Transparenz als Grundlage:

- Kosten sind direkt zuordenbar zu Teams, Services oder Projekten.



Reifegradmodell Cloud-Kostenmanagement

01 Transparenz

Wir sehen die Kosten!

- Toolgestützte Transparenz (z. B. mit IBM Cloudability)
- Kosten sichtbar & zuordenbar
- Reporting & Dashboards

02 Steuerung

Wir entscheiden bewusst!

- Cloud-Kosten als Investitionen
- Klare Produktverantwortung
- Priorisierung statt pauschaler Optimierung

03 Business Value

Wir steuern Wertbeiträge!

- Verknüpfung von Kosten & Business-Zielen
- Governance & FinOps als Betriebsmodell
- Messbarer Beitrag zum Geschäftserfolg

02 Steuerung

➤ **Cloud-Kosten als Investitionen denken**



Jede Skalierung ist eine bewusste Budgetentscheidung.

➤ **Ownership klar zuordnen**



Kosten gehören Produkten, nicht Plattformen

➤ **Priorisieren statt global optimieren**



Nicht nur sparen, sondern gezielt investieren

➤ **Regelmäßige Kosten-Reviews als Entscheidungsformate**

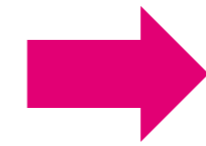
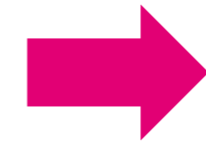


Nicht Reporting, sondern Steuerungsdialoge

02 Steuerung

Beispiele aus der Praxis

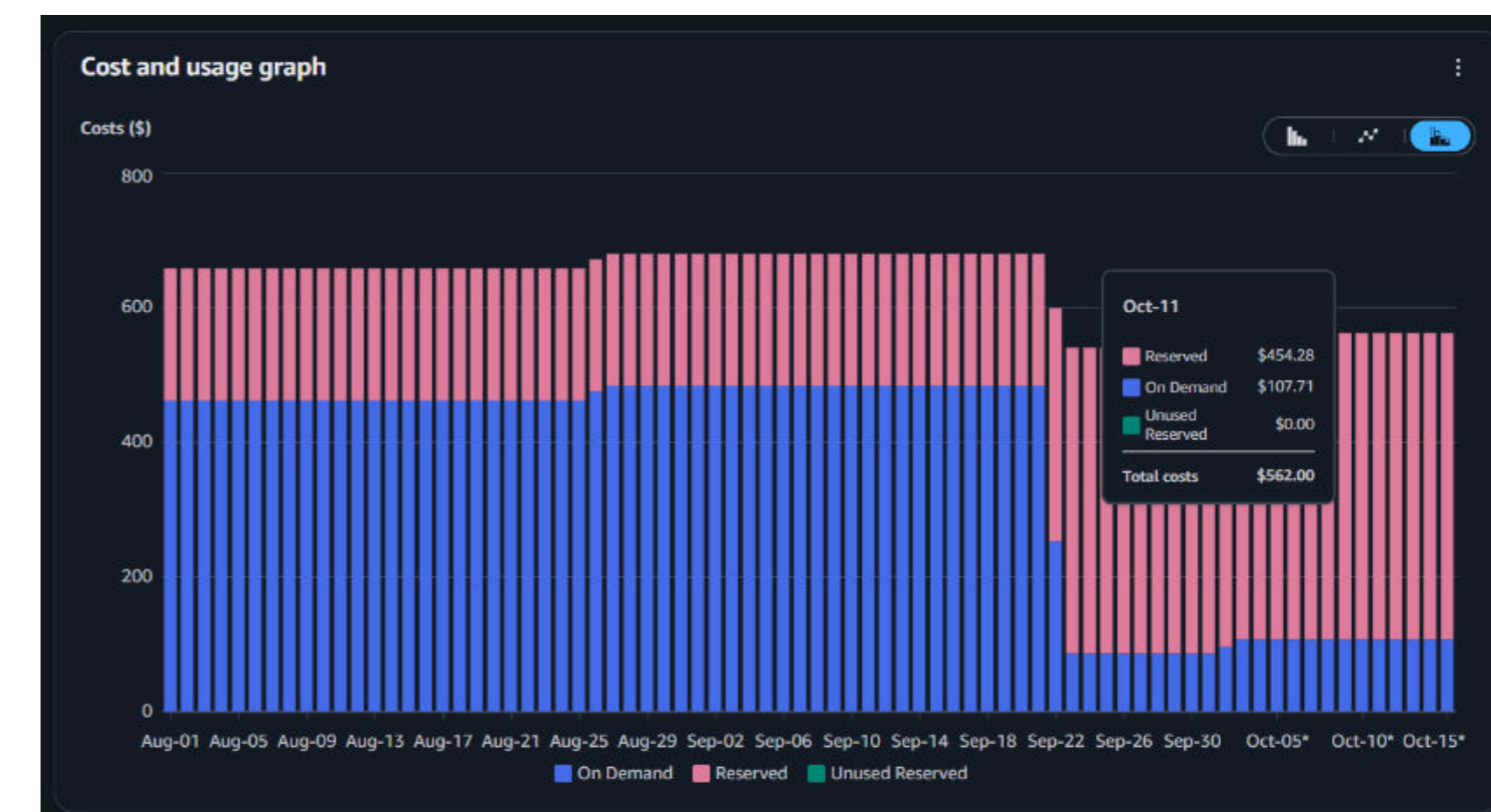
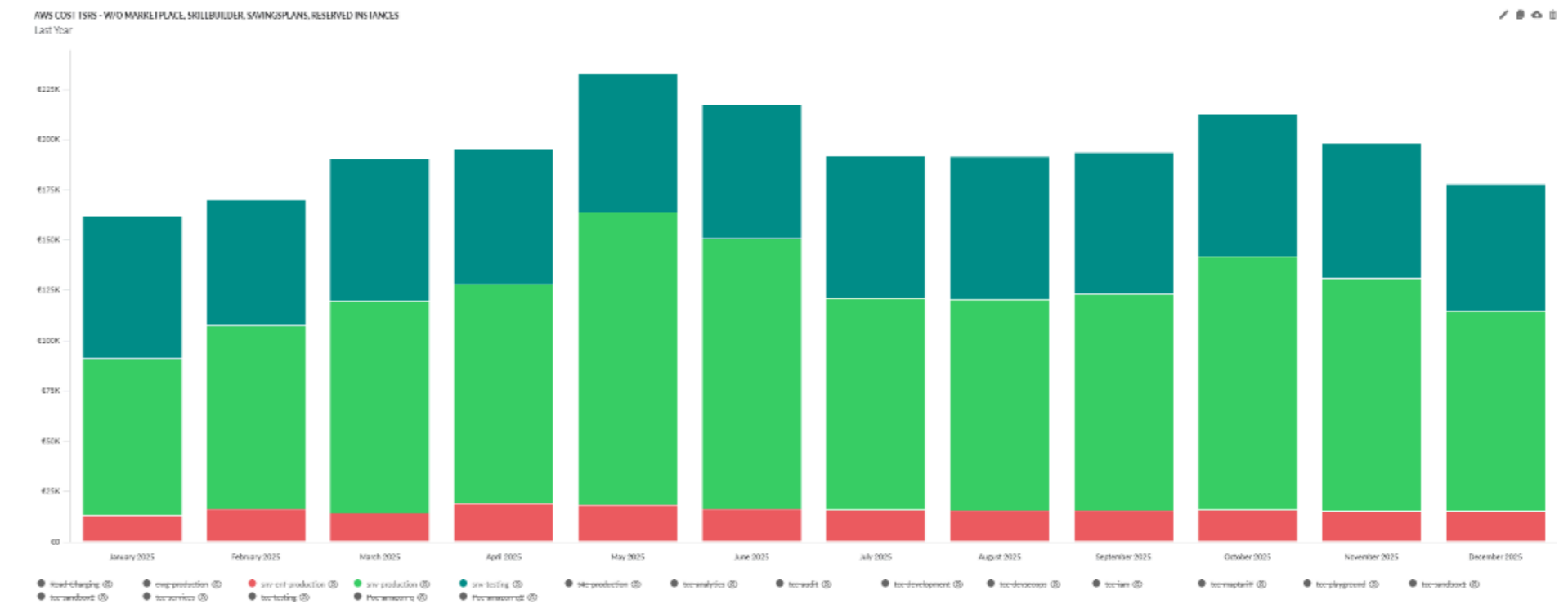
- Produkt-/Systembezogene Kostenverantwortung, d.h. Cloud-Kosten werden pro Produkt sichtbar *und* verantwortet
- Management-Entscheidungen auf Basis von Cloudability-Daten → Kostenentwicklung fließt aktiv in Priorisierungsentscheidungen ein
- Bewusste Trade-offs statt automatischer Optimierung → Performance, Resilienz oder Kosten – je nach Business-Ziel
- Kosten als Teil von Architektur- & Roadmap-Entscheidungen → Architektur kostenbewusst, nicht zufällig



COST OVERVIEW APPLICATION ID
This Month

Pivot By: None

Month (Year)	application-group	application-id	Cost TSI-RS [EUR]
2026-02	casall	edbb	\$4,890.52
2026-02	report	omrs	\$4,268.11
2026-02	mb	mbcb	\$2,709.25
2026-02	mb	mbmb	\$1,818.70
2026-02	tarifall	gils	\$1,664.93
2026-02	casall	eiss	\$1,336.66
2026-02	drsdss	drdb	\$1,267.82
2026-02	report	miss	\$1,231.35
2026-02	tmsall	tmdb	\$1,180.19



Reifegradmodell Cloud-Kostenmanagement

01 Transparenz

Wir sehen die Kosten!

- Toolgestützte Transparenz (z. B. mit IBM Cloudability)
- Kosten sichtbar & zuordenbar
- Reporting & Dashboards

02 Steuerung

Wir entscheiden bewusst!

- Cloud-Kosten als Investitionen
- Klare Produktverantwortung
- Priorisierung statt pauschaler Optimierung

03 Business Value

Wir steuern Wertbeiträge!

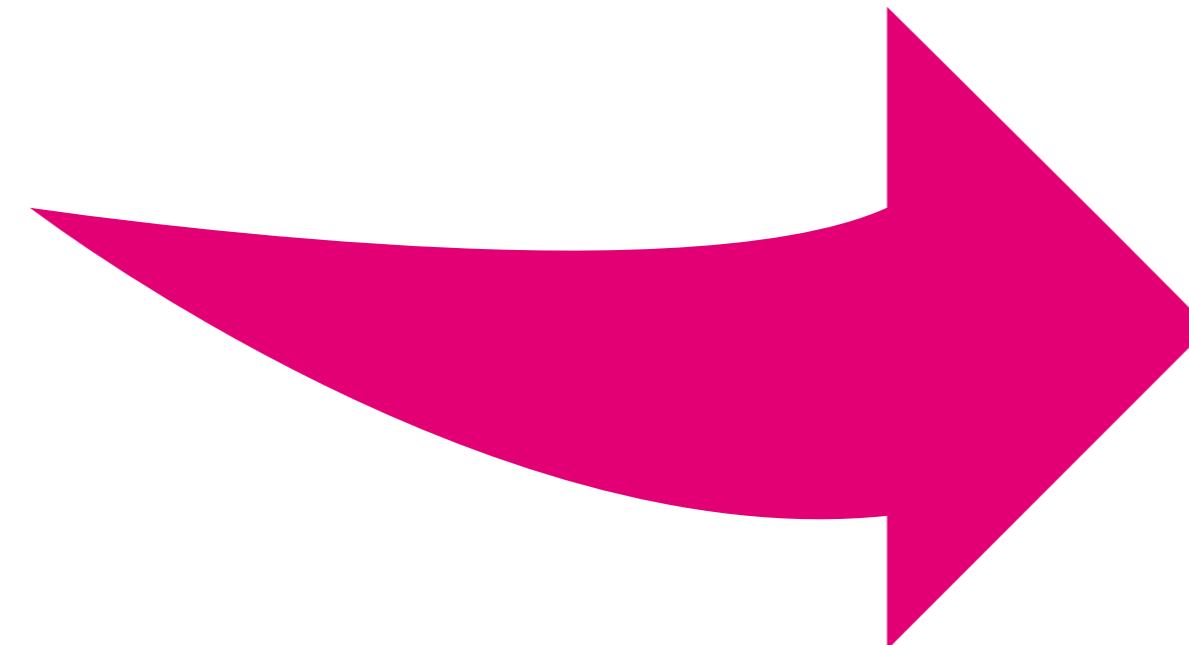
- Verknüpfung von Kosten & Business-Zielen
- Governance & FinOps als Betriebsmodell
- Messbarer Beitrag zum Geschäftserfolg

03 Business Value

Governance & FinOps als Betriebsmodell

- Eindeutige Kostenverantwortung pro Produkt/System
- Feste Entscheidungsformate
- Architektur- & Kostenprinzipien

Klare Regeln schaffen wiederholbare Entscheidungen!



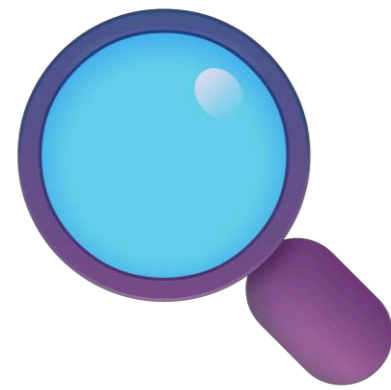
Strukturierte Entscheidungen statt Einzelmaßnahmen

Messbarer Business Value

- Priorisierte Investitionen statt pauschaler Einsparungen
- Bewusste Trade-offs (Kosten vs. Geschwindigkeit vs. Resilienz)
- Transparenter Beitrag zum Geschäftserfolg

Ohne Governance bleibt Business Value Zufall!

Key Takeaways



**Sichtbarkeit ersetzt
keine Entscheidung.**

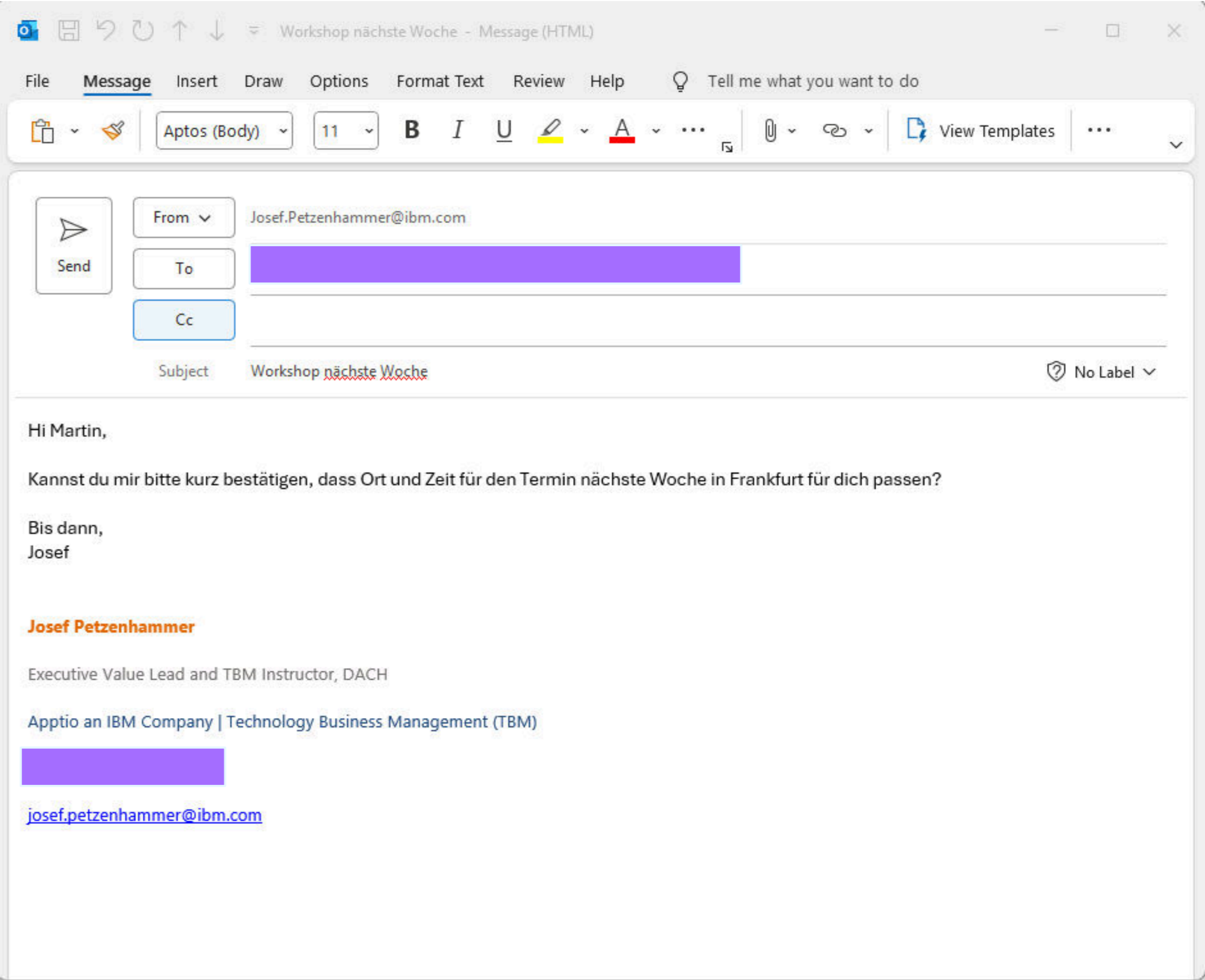
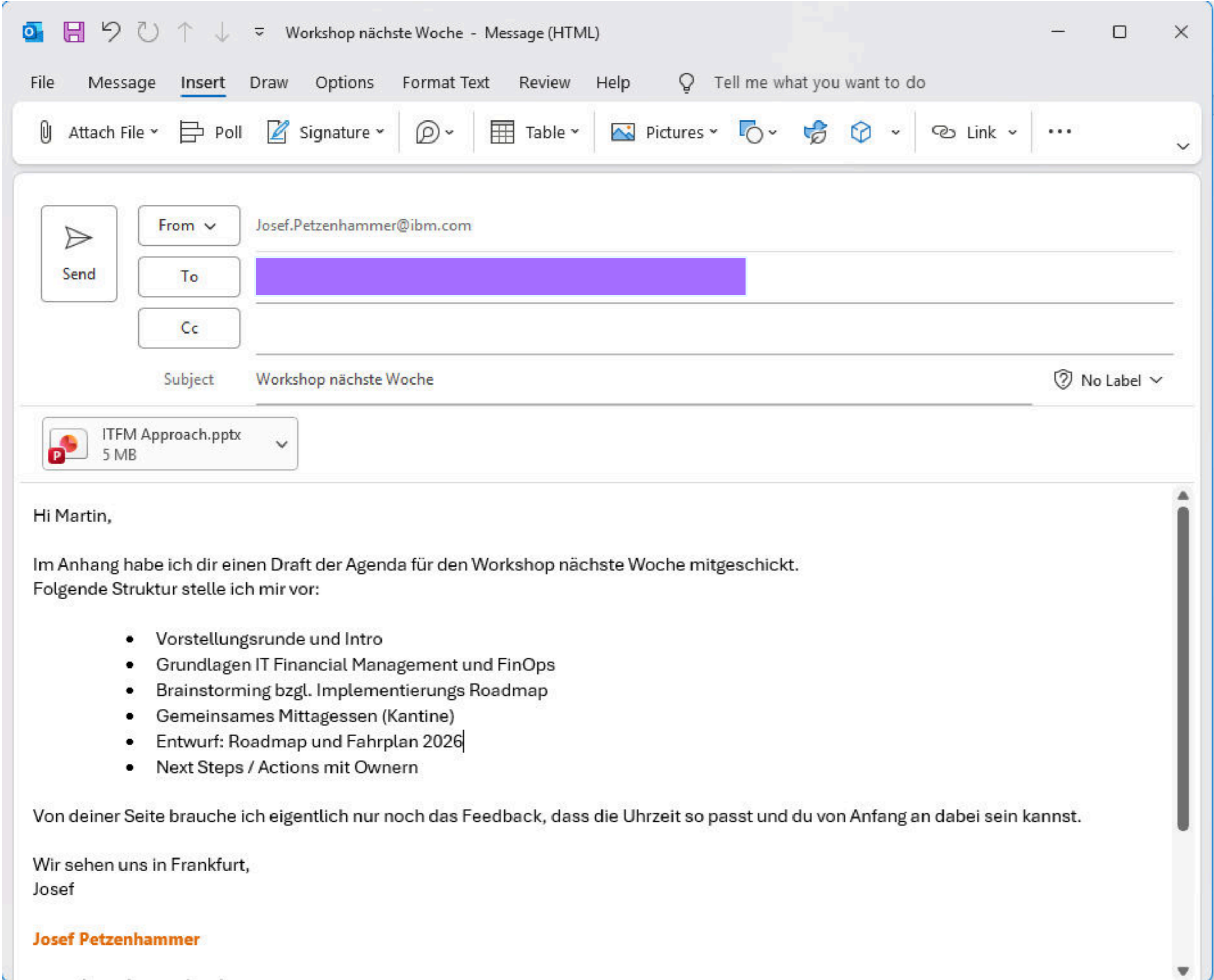


**In der Cloud ist jede
Architektur-
entscheidung eine
Kostenentscheidung.**



**Business Value
entsteht durch bessere
Entscheidungen – nicht
durch Sparen.**

Das AI Zeitalter – Der “Token Wild West”



Kommt uns das nicht bekannt vor?

FinOps und AI – die Parallelen

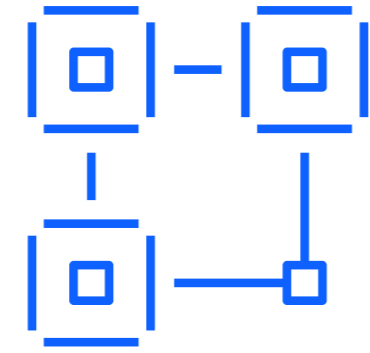


End User Prompts: Der "Viel hilft viel"-Irrglaube

Die Art und Weise des Promptings beeinflusst maßgeblich die Input- und Output-Tokens und damit die Kosten. Die „unendliche Ressource“.

Risiko: Unnötiger Verbrauch von Input-Tokens (die oft teurer sind als Output-Tokens bei großen Kontexten)

Wie war das bei FinOps? Finanzielles Monitoring, Rightsizing, "Quotas" und Schulungen, um (z.B. QA-/Test-) Instanzen dem Nutzen entsprechend zu sizen.

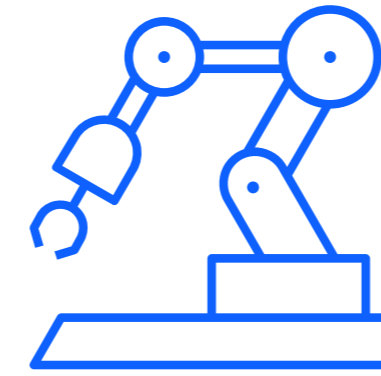


Routinen und Automatisierung: Die Feedback-Schleife

KI wird in ETL-Strecken oder CRM-Automatisierungen eingebaut. Die Kosten skalieren linear mit dem Token-/Datenvolumen.

Risiko: Automatisierung von Tokennutzung, ohne Quality oder Relevance Check

Wie war das bei FinOps? Automatische Skalierung für Datenbanken erforderte maximum burst limits. Finanziell verbundene „Kill Switches“, "Circuit Breaker" und Threshold Monitoring.

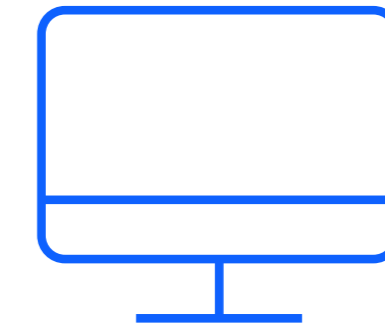


Agentic AI: Der unkontrollierte Dialog

Agenten, die sich gegenseitig Aufgaben delegieren (z.B. Agent A schreibt Code, Agent B prüft ihn, Agent A korrigiert).

Risiko: "Token-Spinning": Agentic Input-Output Loops ohne finanzielle Governance.

Wie war das bei FinOps? Autoscaling-Kaskaden und rekursive Cloud-Funktionen. Strenges Monitoring der "Chain-of-Thought"-Länge und maximale Iterationslimits pro Task.



AI in der Software-Entwicklung

„Vibe-Coding“ erhöht die Zugänglichkeit und Geschwindigkeit der Software-Entwicklung. Demokratisierung von Technologie.

Risiko: Ineffizienzen und Qualitätsmängel in der Architektur und in der Sicherheit von Produktionsinstanzen.

Wie war das bei FinOps? Zugänglichkeit von Ressourcen führte zu Wildwuchs an Cloud-Ressourcen, die niemand überwachte, was zu explodierenden Rechnungen führte.

Kommt uns das nicht
bekannt vor?
FinOps und AI – die
Parallelen

Dezentralisierung

1. Demokratisierung von Technologie
2. Zugänglichkeit
3. Einfachheit (vermeintlich)

Variabilität

1. Intensitätsabhängige Kosten
2. Modellabhängige Kosten
3. Komplexitätsabhängige Kosten

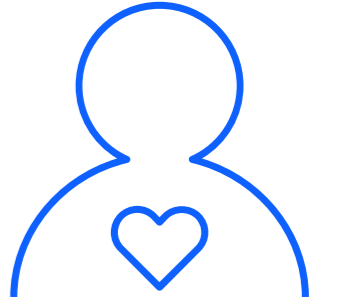
Transparenz- mangel

1. Operationale vs. Finanzielle Transparenz
2. Agentic AI
3. Built-In AI

Kommt uns das nicht
bekannt vor?
FinOps und AI – die
Parallelen

Dezentralisierung

- Bewusstsein durch Transparenz
beim Nutzer
 - Kosten
 - Nutzung
 - Optionen



Variabilität

1. Intensitätsabhängige Kosten
2. Modellabhängige Kosten
3. Komplexitätsabhängige Kosten

Transparenz- mangel

1. Operationale vs. Finanzielle Transparenz
2. Agentic AI
3. Built-In AI

Kommt uns das nicht
bekannt vor?
FinOps und AI – die
Parallelen

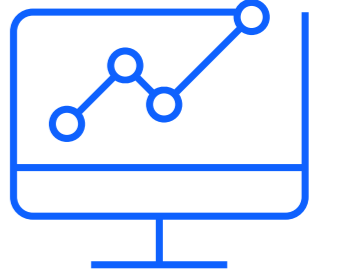
Dezentralisierung

- Bewusstsein durch Transparenz beim Nutzer
 - Kosten
 - Nutzung
 - Optionen



Variabilität

- Kosten pro Modell
- Kosten pro Prompt / Use Case
- Bewertung und Rationalisierung von Alternativen



Transparenz- mangel

1. Operationale vs. Finanzielle Transparenz
2. Agentic AI
3. Built-In AI

Kommt uns das nicht
bekannt vor?
FinOps und AI – die
Parallelen

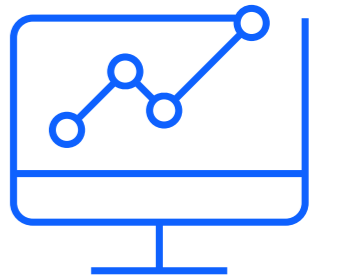
Dezentralisierung

- Bewusstsein durch Transparenz beim Nutzer
 - Kosten
 - Nutzung
 - Optionen



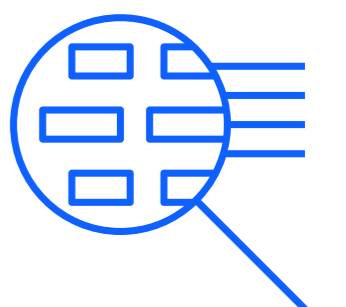
Variabilität

- Kosten pro Modell
- Kosten pro Prompt / Use Case
- Bewertung und Rationalisierung von Alternativen



Transparenz- mangel

- Operationale Transparenz mit Finanzieller Transparenz verbinden
- Kosten/Nutzen-Perspektive für User, Applikationen, Produkte, Geschäftsbereiche...



Ganzheitlicher Ansatz zur Optimierung des AI-Mehrwerts

Strategie

CIO



Herausforderung:

Entscheidungen ohne Fokus auf Ressourcen-Impact. Erforderlich: Portfolio-Gesamtblick auf Prioritäten, Budget und ROI.

Kernproblem:

Für 67 % der CIOs steigt der Rechtfertigungsdruck für KI-Investitionen schneller als die Messbarkeit des ROI.

Kosten

IT Finance



Herausforderung:

Ausgaben sind verteilt und schwer nachverfolgbar. Erforderlich: Einheitlicher AI TCO über Cloud, Vendoren und Personal hinweg.

Kernproblem:

Nur 7 % der KI-Ausgaben werden über Innovationsbudgets erfasst.

Nutzung

Cloud CoE



Herausforderung:

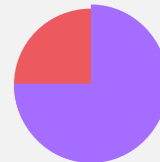
Allokation oder Rightsizing von KI-Workloads. Benötigt Telemetrie + Tags, um GPU-Ausgaben Projekten zuzuordnen.

Kernproblem:

37 % der Unternehmen nutzen bereits 5+ LLMs in Produktion

Mehrwert

Innovation or AI Lead



Herausforderung:

Fehlende Wertschöpfung oder Benchmarks. Benötigt Tools, um den KI-Nutzen nachzuverfolgen und zu belegen.

Kernproblem:

Die KI-Ausgaben sind von 2023 bis 2025 um das Dreifache gestiegen, doch den meisten Organisationen fehlt Governance

1

AI-ROI-Messung

Verknüpfung von KI-Investitionen mit dem Geschäftswert durch TBM- und FinOps-Frameworks.

2

Financial Predictability

Forecasting von AI-bezogenen Ausgaben zur Vermeidung von Budgetüberschreitung

3

Maximize AI Spending

Identifizierung von Ineffizienzen bei der KI-Ressourcenallokation für kosteneffiziente Investitionen.

4

AI-Investment-Transparenz

Nachverfolgung der KI-Gesamtkosten über Cloud-, Personal-, Infrastruktur- und Anwendungsebene hinweg.

AI für Apptio: KI-gestützte TBM und FinOps-Prozesse

- Aufzeigen zentraler Kostentreiber und Optimierungspotenziale.
- AI-Forecast und Planung
- Automatisierung von Finanzaufgaben
- Unterstützung der Anwender durch KI-Assistenten und kontextbezogene Empfehlungen



Apptio für AI: Messen, Managen und Optimieren von AI Investments

- Volle Transparenz über die KI-Gesamtkosten (TCO) für Cloud, Personal, Tools und Anbieter.
- GPU-, Token- und API-Nutzung nach Projekt, Team und Umgebung.
- Allokation von KI-Ausgaben an Geschäftsbereiche und Use-Cases.
- Cost/Value-Messung von GenAI-Modellen

KI-Governance, Ethik und Datenmanagement für Unternehmen – ergänzt und ermöglicht durch das vollständige IBM watsonx Portfolio.

AI ist Realität

- Klare Transparenz über Nutzung, Budgets und Mehrwerte ist erforderlich.
- “Testballons” ohne messbaren Mehrwert sind zu teuer.

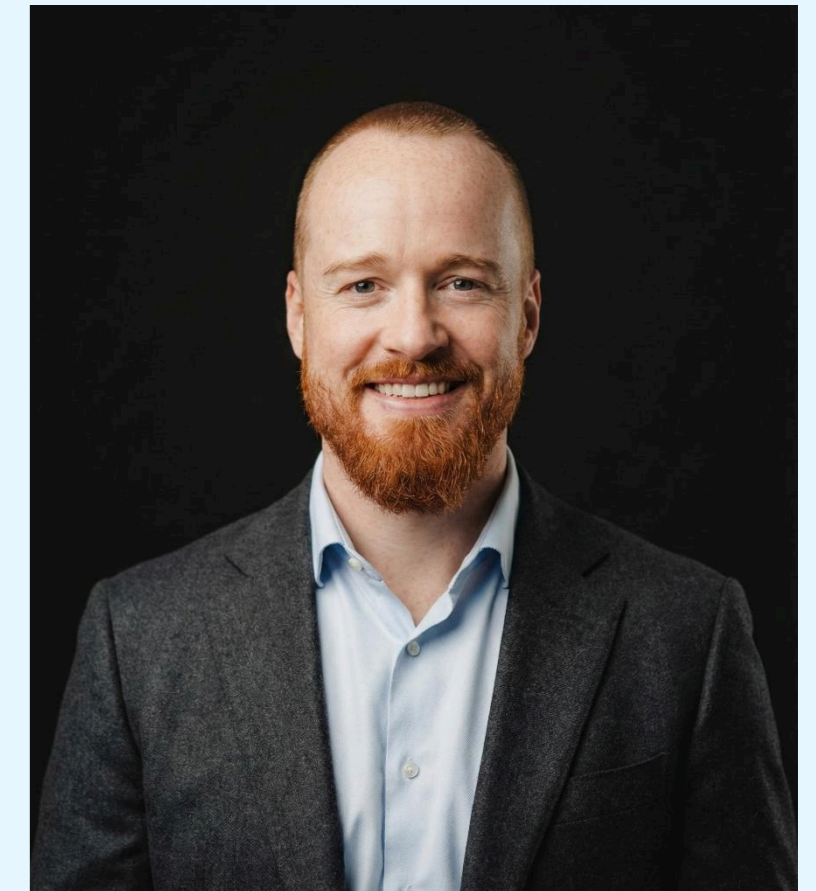
ITFM für AI ist nicht neu

- TBM und FinOps liefern bereits die wichtigsten Grundlagen.
- Zusammenspiel zwischen IT, Finanz und Business ist “Key”.

Nutzer und Nutzung

- Die gesamte Wertschöpfungskette von der Technologie zum Nutzer muss modelliert werden.

Vielen Dank für Ihre Aufmerksamkeit!



Josef Petzenhammer
Executive Value Lead (EMEA)
Apptio, an IBM Company