

Technology Arts Sciences TH Köln

Technische Hochschule Köln

Fakultät für Informatik und Ingenieurwissenschaften

BACHELORARBEIT

Kostenüberwachung und -optimierung für Cloud-Dienste am Beispiel von Amazon Web Services

Vorgelegt an der TH Köln Campus Gummersbach
im Studiengang Wirtschaftsinformatik

ausgearbeitet von:

CARLO MENJIVAR 11117929

Erstprüfer: Prof. Dr. Roman Majewski

Zweitprüfer: Thomas Raser

Gummersbach, 1x Feb 2021

Abstract

Zusammenfassung

In dieser Arbeit werden Werkzeuge untersucht, die einen klareren Überblick über die finanziellen Ressourcen schaffen[DAS MUSS GEÄNDERT WERDEN]. Mit den gesammelten Informationen dienen sie dazu, direkte Maßnahmen zu ergreifen. Darüber hinaus werden allgemeine Optimierungsmaßnahmen aufgezeigt, die bereits über die Jahre hinweg von anderen Nutzern getestet wurden und von Amazon Web Services (als Best Practices) empfohlen werden. Die Grundlage dieser Recherche sind Empfehlungen von Cloud-Anbietern bezüglich Kostenüberwachung und -optimierung, Erfahrungen von Experten in dem Fachgebiet und aktuelle Fachliteratur.

Es ist besonders interessant für Teams, die Cloud-Dienste in aktuellen Projekten nutzen und ihre Kosten in der Cloud besser verstehen und optimieren wollen. Wenn die Kosten für Cloud-Dienste wie alle anderen Kosten betrachtet werden, ist es konsequent, über ihre Kontrolle und Optimierung nachzudenken. Ein häufiges Problem ist, dass Kosten entstehen, die sich der Kontrolle der Nutzer entziehen. Aus diesem Grund stehen Unternehmen die bereits On-Premise IT-Infrastruktur nutzen, einem Wechsel kritisch gegenüber, obwohl ihnen die Flexibilität von Cloud-Diensten bessere Wettbewerbsvorteile bieten würde. Deshalb sind die in dieser Arbeit aufgezeigten Werkzeuge und Maßnahmen relevant für diejenigen, die von einem Wechsel von klassischen Modellen, bekannt als On-Premise, zu Cloud basierten Modellen profitieren möchten.

Abstract

Platz für das englische Abstract....

Inhaltsverzeichnis

| | |
|--|-----------|
| Abstract | 1 |
| Abbildungsverzeichnis | 5 |
| Abkürzungsverzeichnis | 6 |
| 1 Einleitung | 7 |
| 1.1 Motivation | 7 |
| 1.2 Problemstellung | 7 |
| 1.3 Zielsetzung | 8 |
| 1.4 Struktur der Arbeit | 8 |
| 2 Grundlagen | 10 |
| 2.1 Cloud Economics | 10 |
| 2.1.1 Skalierbarkeit | 10 |
| 2.1.2 Flexibilität | 11 |
| 2.1.3 Selbstbedienung | 12 |
| 2.1.4 Keine Vorabkosten | 12 |
| 2.1.5 Technische Fachkompetenz | 12 |
| 2.2 Amazon Cloud-Dienste | 13 |
| 3 Zahlungsmodelle | 15 |
| 3.1 On-Demand-Instanzen | 15 |
| 3.2 Reservierte Instanzen und Saving Plans | 16 |
| 3.3 Spot Instanzen | 18 |
| 3.4 Amazon EC2 Fleet[rev] | 19 |
| 3.5 Anwendungsfall: TrueCar[rev] | 20 |
| 4 Kostenüberwachung | 25 |
| 4.1 AWS CloudWatch | 26 |
| 4.2 AWS Cost-Explorer | 28 |
| 4.3 AWS Trusted Advisor | 29 |
| 4.4 Überwachungswerkzeuge gemäß ihrer Verwendung | 31 |
| 5 Optimierungsmaßnahmen | 33 |
| 5.1 EC2 Automatische Skalierung | 33 |

| | | |
|---|---|-----------|
| 5.1.1 | Zeitgesteuerte Skalierung | 34 |
| 5.1.2 | Dynamisches Auto Scaling | 35 |
| 5.1.3 | Manual Scaling | 35 |
| 5.1.4 | Predective Scaling | 35 |
| 5.2 | S3 Optimierung | 36 |
| 5.2.1 | Die richtige Speicherklassen wählen | 36 |
| 5.2.2 | Lebenszyklus-Konfiguration | 37 |
| 5.2.3 | Intelligent-Tiering | 39 |
| Zusammenfassung und Ausblick | | 41 |
| | Bewusstsein in der gesamten Organisation | 41 |
| | Die richtige Personen finden, Owneship verbreiten | 41 |
| | 5G is comming | 42 |
| | Rentabilität bei der Optimierungsmaßnahmen | 42 |
| Glossar | | 43 |
| Quellenverzeichnis | | 45 |
| A Anhang | | 51 |
| A.1 | Anhang X | 51 |
| Erklärung über die selbständige Abfassung der Arbeit | | 52 |

Abbildungsverzeichnis

| | | |
|----|--|----|
| 1 | 2020 überholt die Cloud lokale Speichermedien | 14 |
| 2 | On-Demand Preise für Amazon EC2 | 16 |
| 3 | Mögliche Einsparungen bei Reserved Instances and Saving Plans laut AWS | 17 |
| 4 | Mögliche Einsparungen durch Vorauszahlungen | 18 |
| 5 | Monatliche Kosten für eine On-Demand-Instanz im Vergleich zu einer re- servierten Instanz | 22 |
| 6 | Vergleich der Zahlungsmodelle | 23 |
| 7 | Trennung der Kosten durch Tags | 25 |
| 8 | AWS Trusted Advisor Kategorien | 29 |
| 9 | Überwachungswerkzeuge gemäß ihrer Verwendung | 31 |
| 10 | Ungenutzte Rechenkapazität ohne automatische Skalierung | 33 |
| 11 | Berechnung für ein nicht produktives Umgebung mit Zeitgesteuerte Ska- lierung | 34 |
| 12 | Nutzung von Tinder, OkCupid und Netflix pro Stunde | 36 |
| 13 | Kostenvergleich durch Nutzung von unterschiedlichen Speicherklassen . . . | 39 |
| 14 | Funktionsweise von Intelligent-Tiering | 40 |

Abkürzungsverzeichnis

AWS Amazon Web Services

API Application Programming Interface

CI/CD Continuous Integration / Continuous Deployment

TCO Total Cost of Ownership

EC Elastic Compute

PAYG Pay-as-you-go

ASG Auto Scaling Group

1 Einleitung

1.1 Motivation

Die zunehmende Digitalisierung von Geschäftsmodellen, die auch durch die Corona-Pandemie vorangetrieben wird[45], lässt Cloud-basierte Applikationen an Bedeutung gewinnen[44]. Als direkte Folge davon steigt die Nachfrage nach Server- und Speicherkapazität. Amazon Web Services, kurz AWS, wurden unter anderem als Fallbeispiel für diese Arbeit ausgewählt wegen seiner frühen Präsenz (2006) als Cloudanbieter und seines großen Angebotes an Dienstleistungen, welche für zahlreiche Anwendungsfälle geeignet sind. Eine Recherche von Gartner positioniert AWS als Marktführer in der Magic Quadrant¹ für Cloud-Infrastruktur und Plattform-Services 2021[34].

Kostenoptimierung für Cloud-Dienste ist ein wichtiger Punkt, da man ohne Optimierungsmaßnahmen mit höheren Kosten rechnen muss als bei On-Premise Systemen.

”Indeed, if you run the cloud the same way you run your on-premise data center, you are almost certain to incur higher expenses. It is necessary to use the following key cloud cost optimization techniques in order to successfully save money on the cloud.”²

1.2 Problemstellung

Die Verwendung von Cloud-Diensten bringt viele Vorteile mit sich. Zum Beispiel kurzfristige Erhöhung oder Verringerung der Speicher- und Rechenkapazität, sowie Zugriff auf unterschiedliche Speicherarten, die genau an individuelle Anwendungsfälle angepasst sind. All diese Lösungen sind in wenigen Minuten einsatzfertig.

In einer Umfrage haben circa 50% der Unternehmen die Verwaltung der Kosten für den Betrieb von Cloud-Workloads als großes Hindernis genannt. Mehr als die Hälfte der Befragten haben geäußert, dass sie Schwierigkeiten haben, alle Kosten für Cloud-Workloads zu rechtfertigen.

¹Der Magic Quadrant ist eine zweidimensionale Matrix mit vier Quadranten. Jeder Quadrant steht für einen Unternehmenstypus im Markt. Im Uhrzeigersinn von links unten beginnend sind dies: Niche Players, Challengers, Leader und Visionaries.

²[1], Seite 152

„In its Stratecast Predictions 2018, Frost & Sullivan noted that 53% of IT leaders surveyed cited “managing costs to run cloud workloads” as a huge obstacle, and over 50% have difficulty justifying the expenses of some public cloud workloads.“ [40]

Diese Bachelorarbeit beschäftigt sich mit ebendieser Problematik, um herauszufinden, wie Unternehmen mit den passenden Werkzeugen die Kosten ihrer Cloud-Dienste überwachen und im Blick behalten können. Zum Beispiel können frühzeitige Benachrichtigungen alarmieren, wenn Cloud-Dienste mehr Kosten verursachen als geplant.

Außerdem sollte untersucht werden, wie mit der richtigen Auswahl an Diensten Kosten optimiert werden. In dieser Arbeit wird versucht zu beantworten, wie Kosten bei Cloud-Diensten überwacht werden können. Auf Grundlage dieser Information werden Optimierungsmöglichkeiten gesucht. Es wird untersucht, welche Maßnahmen nötig sind, um unerwartet hohe Kosten bei Cloud-Diensten zu vermeiden. Darüber hinaus werden Empfehlungen von Cloud-Experten berücksichtigt, um Kosten von Cloud-Diensten zu minimieren. Diese Arbeit untersucht speziell S3-Speichereinheiten und EC2-Server-Instanzen.

1.3 Zielsetzung

Die vorliegende Arbeit betrachtet die von AWS angebotenen Überwachungswerkzeuge, um ein tiefergehendes Verständnis der Entstehung von Kosten durch die Nutzung von Cloud-Diensten zu gewährleisten. Mit den von AWS zur Verfügung gestellten Maßnahmen sollen die Nutzung und damit die Kosten von Cloud-Diensten reduziert werden.

1.4 Struktur der Arbeit

Diese Bachelorarbeit ist in folgende Kapitel unterteilt:

Kapitel 2 befasst sich mit dem Begriff Cloud-Economy und erläutert das Potenzial der Cloud-Diensten im wirtschaftlichen Sinne. Die Cloud-Dienste EC2-Instanzen und S3 Speichereinheiten werden ebenfalls kurz erklärt.

Kapitel 3 zeigt die verschiedenen Zahlungsmodelle für EC2-Instanzen. Es werden Kriterien vorgestellt, die helfen sollen, sich für das richtige Zahlungsmodell bei verschiedenen Szenarien zu entscheiden.

In Kapitel 4 werden die Werkzeuge eingeführt, die zur Überwachung der Kosten von Cloud-Diensten eingesetzt werden.

Kapitel 5 befasst sich mit Optimierungsmaßnahmen insbesondere für EC2-Instanzen und S3 Speichereinheiten.

2 Grundlagen

In diesem Grundlagenkapitel werden Erfolgchancen für Unternehmen aufgelistet, die Cloud-Dienste in ihre Geschäftsprozesse integrieren. Mit Cloud-Diensten sind die Dienste eines beliebigen Cloud-Anbieters im Allgemeinen gemeint und nicht ausschließlich Amazon Web Services(AWS-Dienste). Es wird ebenfalls erklärt warum Kostenoptimierung und -überwachung relevant für Unternehmen sind.

Folgende Ergebnisse könnten durch die Einführung von Überwachungs- und Optimierungsmaßnahmen erreicht werden:

- Die Möglichkeit, die Kosten verschiedener Projekte, die über dieselbe Infrastruktur laufen, zu trennen. Auf diese Weise kann zwischen Projekten, die mehr, und Projekten, die weniger Kosten verursachen unterschieden werden.
- Eine beachtliche Erhöhung der finanziellen Rentabilität im Unternehmen.
- Eine geringere Ungewissheit bei der Umsetzung von cloudbasierten Systemen.
- Mehr Kontrolle über die Gesamtkosten des Betriebs (TCO³)⁴.

2.1 Cloud Economics

Cloud Economics untersucht die Kosten und die Vorteile von Cloud Computing und die, der dahinterstehenden wirtschaftlichen Grundsätze. Das On-Demand Prinzip, besitzt die Flexibilität, die Rechenkapazität je nach Bedarf anzupassen. Es entfällt die Notwendigkeit, hohe Investitionen in Hardware zu tätigen, wie bei On-Premise-Systemen. Durch den Verzicht auf Hardware entfallen die Kosten für Reparatur und Wartung. Cloud-Anbieter übernehmen viele Verwaltungsaufgaben. Das führt zu einer Abnahme der nötigen Fachkraft[33]. Die Nutzung von Cloud-Diensten ist in unabhängiger Weise möglich; in Selbstbedienung und mit der Freiheit Dienste ohne Einschränkungen zu nutzen. Das bedeutet jedoch gleichzeitig, dass Nutzer von Cloud-Diensten Verantwortung für die anfallenden Kosten übernimmt.

2.1.1 Skalierbarkeit

Skalierbarkeit bezieht sich in dieser Arbeit auf die Möglichkeit, die Kapazität von Cloud-Diensten zu skalieren. Um die Leistung der IT-Infrastruktur aufrecht zu halten, ist es zum

³TCO: Total Cost if Ownership

⁴[32],Ubuntu, delivered by Canonical:A business guide to hybrid/multi-cloud, Seite 2

Beispiel möglich, das Serversystem so zu konfigurieren, dass es auf wechselnde Lastanforderungen reagiert. Auf diese Weise kann Zeit mit der Verwaltung von IT-Infrastruktur gespart werden, welche dann genutzt werden kann, um sich auf die wesentlichen Geschäftsaktivitäten zu konzentrieren.⁵

Dies war der Fall bei Walgreens 2020 in den Vereinigten Staaten. Sie haben unter anderem 750 virtuelle Maschinen und SAP HANA auf Azure Instanzen migriert.

„By getting out of the business of managing datacenters, WBA[Walgreens Boots Alliance] can spend less time worrying about managing IT resources and more time focusing on what it’s really good at—delivering great health-care and retail experiences to its customers. Azure also gives WBA an opportunity to better utilize the capabilities of its SAP implementation. “One of the key reasons for moving to Azure was so that we could take advantage of the scalability that SAP HANA is capable of,” explains Regalado. “Instead of using extremely big SAP HANA Large Instances, we can start using smaller VMs[virtuelle Maschinen] and then scale out.,”⁶

2.1.2 Flexibilität

Mit Flexibilität ist gemeint, die Möglichkeit Cloud-Dienste, wenn nötig, in Auftrag zu geben und zu kündigen, wenn sie nicht mehr benötigt werden. Das unter den mit dem Cloud-Anbieter vereinbarten Bedingungen. Für Cloud-Dienste gibt es im Allgemeinen eine Vielzahl von Optionen, von denen einige Beispiele unten aufgeführt sind:

- Verschiedene Betriebssysteme, ohne oder mit Lizenzierung.
- Die meistverbreiteten Programmiersprachen, unter anderem Java, C++, Go, JavaScript und Python.[5]
- Hosting für statische Webseiten und Webanwendungen[6].
- Populäre relationale und nicht relationale Datenbanken[11].
- Vielfältige Hardware-Konfigurationen.

⁵[2], WS Certified Solutions Architect - Associate (SAA-C02), Seite 29

⁶[28] Microsoft Customer Story-Walgreens Boots Alliance delivers superior customer service with SAP solutions on Azure

Durch die Vielzahl der verfügbaren Diensten ist es möglich, Prototypen und Experimente in kurzer Zeit durchzuführen⁷. Softwareprojekte können schnell auf den Markt gebracht werden. Je nach ihrem Erfolg ist es möglich, sinnvolle Entscheidungen zu treffen. Wenn ein Projekt, aus welchen Gründen auch immer, kurzfristig eingestellt werden muss, könnten alle damit verbundenen Kosten ausfallen. Denn im Gegensatz zu On-Premise-Infrastrukturen gibt es keine Bindung an kostspielige Hardware.

2.1.3 Selbstbedienung

Mit geringem Aufwand ist es möglich, Cloud-Dienste eigenständig einzurichten. Dies hat den Vorteil, dass keine weiteren Personen wie externe Spezialisten oder die Vertriebsabteilung des Cloud-Anbieters benötigt werden⁸. Andererseits besteht die Gefahr, dass hohe ungewollte Kosten entstehen, wenn jemand versehentlich oder in unverantwortlicher Weise Dienstleistungen in Anspruch nimmt.

2.1.4 Keine Vorabkosten

Das Pay-as-you-go-Modell(PAYG) wird von einer Reihe von Cloud-Anbietern angeboten. Dies erfordert keine Vorauszahlungen für die Nutzung von Cloud-Diensten. Wenn nur für die monatlich verbrauchten Diensten bezahlt wird, verringert sich die Anfangsinvestition in die IT-Infrastruktur oder fällt ganz weg. Dies ist besonders für kleine Unternehmen interessant, die nicht über die finanziellen Mittel verfügen, um in eine IT-Infrastruktur zu investieren. Es besteht jedoch die Möglichkeit, bestimmte Beträge für die zu konsumierende Dienste im Voraus zu bezahlen. Im Unterkapitel 3.2 wird eine Berechnung der Einsparungen durch die teilweise oder vollständige Vorauszahlung der Kosten für die Nutzung von Serverinstanzen gezeigt.

2.1.5 Technische Fachkompetenz

Es ist zu bedenken, dass weitere Investitionen wie technische Schulungen für das Personal erforderlich werden. TÜV Rheinland bietet Kurse zur Ausbildung von Cloud Architekten an. Die Kurse dauern drei Tage und kosten 2.136,05 € pro Teilnehmer. Maßnahmen wie die genannten Kurse wirken einem der Hauptprobleme entgegen, mit denen Unternehmen bei der Migration in die Cloud konfrontiert sind. In der von Accenture im Jahr 2020 durchgeführten Umfrage gaben 38% der Befragten an, dass fehlende Kompetenzen im

⁷[33], IDC Business Value of AWS 2015 Seite 7

⁸[1], Cloud Computing Basics: a Non.-Technical Introduction, Seite 28

Unternehmen im Bezug auf die Cloud ein Hindernis für eine Cloud-Migration ist⁹.

2.2 Amazon Cloud-Dienste

Ab dieser Stelle der Arbeit liegt der Fokus auf den Cloud-Diensten von Amazon Web Services, die als AWS-Dienste bezeichnet werden.[Rev] Einer der am häufigsten genutzten AWS-Dienste ist Amazon Elastic Computing Instances EC2, mit dem virtuelle Maschinen erstellt werden können¹⁰. Amazon Elastic Computing EC2-Instanzen werden ab sofort als EC2-Instanzen bezeichnet. Ein weiterer wichtiger Dienst ist Amazon Simple Storage Service (S3), der zum Speichern von Objekten verwendet wird. Deshalb konzentrieren sich in dieser Arbeit die Überwachungs- und Optimierungsmaßnahmen hauptsächlich auf EC2-Instanzen und S3-Speichereinheiten. Wie Lynn Langit, eine erfahrene Cloud-Architektin, feststellt, können bis zu 80% der Rechnung aus Gebühren für EC2-Instanzen bestehen¹¹.

Objekte sind in AWS die Grundeinheit in welchen Dateien in den S3-Speichereinheiten gespeichert werden. Neben den Objekten werden Metadaten, wie das Datum der Objekterstellung und das Datum der letzten Aktualisierung gespeichert. Laut des AWS Solutions Architekten Daniel Peña Silva¹² ist Amazon S3 einer der am häufigsten genutzten AWS-Dienste.

Wie in Abbildung 1 zu sehen ist, werden darüber hinaus seit 2020 weltweit mehr Daten in Serverfarmen als auf lokalen Geräten gespeichert¹³. Dies bietet Vorteile im Bezug auf die Geschwindigkeit der Arbeitsabläufe, birgt aber auch Risiken wie Datendiebstahl. Das Thema Datendiebstahl wird in dieser Arbeit nicht behandelt; da es den Rahmen der Recherche sprengen würde.

Dieses grundlegende Kapitel hat einige potenzielle Vorteile der Nutzung von Cloud-

⁹[1], Accenture Dienstleistungen GmbH. Hohe Erwartungen an die Cloud: Hürden meistern, Mehrwert maximieren, Seite 11

¹⁰[46] Cloud infrastructure services vendor market share worldwide from 4th quarter 2017 to 3rd quarter 2021

¹¹[36]LinkedIn Learning: AWS Controlling Cost by Lynn Langit

¹²[35]LinkedIn: Listado de todos los Servicios de AWS

¹³[43]2020 überholt die Cloud lokale Speichermedien

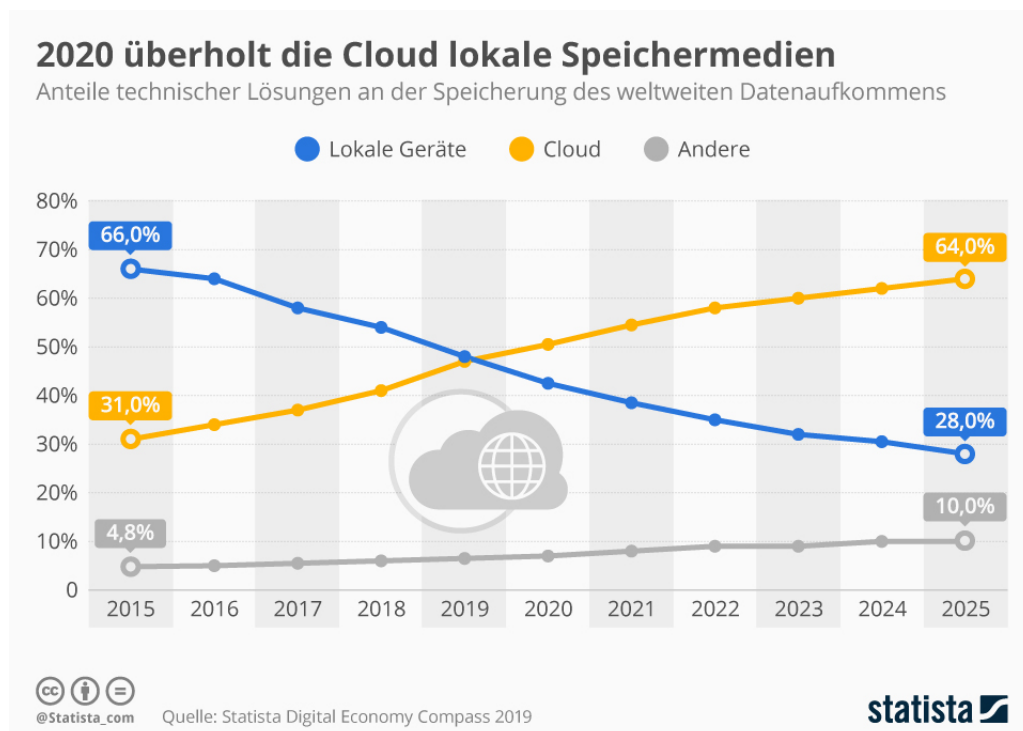


Abbildung 1
2020 überholt die Cloud lokale Speichermedien [43]

Diensten für Unternehmen aufgezeigt. Darüber hinaus geht der Trend in den letzten Jahren zur Nutzung von Cloud-basierten Diensten. Das nächste Kapitel befasst sich mit den Zahlungsmodellen für EC2-Instanzen und den Überlegungen, die bei der Wahl dieser Modelle in verschiedenen Szenarien zu berücksichtigen sind.

3 Zahlungsmodelle

Die Nutzung von EC2-Instanzen ist mit einem Zahlungsmodell verbunden. Die Wahl des Zahlungsmodells ist von entscheidender Bedeutung, um den besten Preis für EC2-Instanzen zu erzielen. Die von Amazon Web Services angebotenen Zahlungsmodelle werden im Folgenden dargestellt.

Das On-Demand-Modell beinhaltet keine langfristigen Verpflichtungen, es ist daher die teuerste Alternative, die auf Stundenbasis berechnet wird. Die Modelle Saving Plans und Reserved Instances erfordern den Abschluss von Verträgen über ein oder drei Jahre, um günstige Preise zu erhalten. EC2-Spot-Instanzen sind das kostengünstigste Modell, sie haben aber den Nachteil, dass ihre Verfügbarkeit nicht immer garantiert ist. Jedes Zahlungsmodell hat seine Vor- und Nachteile und eignet sich für unterschiedliche Anwendungsfälle. Gute Ergebnisse können auch durch die Kombination mehrerer Zahlungsmodelle erzielt werden. Dies wird in Unterkapitel 3.4 behandelt.

In dieser Arbeit wird nicht darauf eingegangen, wie die richtige Server-Instanz ausgewählt werden sollte, da die Auswahl von individuellen Anforderungen abhängt, die von Fall zu Fall unterschiedlich sind. Im Allgemeinen wird empfohlen Instanzen so nahe wie möglich an den Ressourcen, mit denen sie kommunizieren werden, zu platzieren. Die beste Leistung wird außerdem angestrebt, indem sich diese Instanzen in räumlicher Nähe zur Mehrzahl der Endnutzer, befinden.

3.1 On-Demand-Instanzen

Bei diesem Zahlungsmodell besteht keine Notwendigkeit, ein festes Anfangsbudget festzulegen. Die Kosten richten sich nach dem Verbrauch auf der Grundlage der Nutzungsstunden. Dieses Modell eignet sich für Projekte, deren Entwicklung unvorhersehbar ist und die Möglichkeit besteht, dass das es in kurzer Zeit abgeschlossen sein wird, sodass es nicht Sinnvoll ist, eine langfristige Verpflichtung einzugehen.

Die Preise beim dem On-Demand Zahlungsmodell variiert je nach Instanz Typ, Region und der übertragenen Datenmenge. Die aktuellen Preise für die verschiedenen Regionen sind auf der Amazon-Website in der Sektion EC2 - On-Demand-Preise¹⁴ zu finden. In der Abbildung 2 werden die für die Region Ohio verfügbaren Linux-Instanzen gezeigt.

¹⁴[4]AWS On-Demand Instances Pricing

Region, Betriebssystem, Instance-Typ und vCPU auswählen, um Tarife anzuzeigen

Region: Betriebssystem:

Instance-Typ: vCPU:

363 von 363 verfügbaren Instances werden angezeigt

< 1 2 3 4 5 6 7 ... 19 >

| Instance-Name ▲ | On-Demand-Stundensatz ▼ | vCPU ▼ | Arbeitsspeicher ▼ | Speicherung ▼ | Netzwerkleistung ▼ |
|-----------------|-------------------------|--------|-------------------|---------------|--------------------|
| a1.medium | 0,0255 USD | 1 | 2 GiB | Nur EBS | Bis zu 10 Gigabit |
| a1.large | 0,051 USD | 2 | 4 GiB | Nur EBS | Bis zu 10 Gigabit |
| a1.xlarge | 0,102 USD | 4 | 8 GiB | Nur EBS | Bis zu 10 Gigabit |
| a1.2xlarge | 0,204 USD | 8 | 16 GiB | Nur EBS | Bis zu 10 Gigabit |
| a1.4xlarge | 0,408 USD | 16 | 32 GiB | Nur EBS | Bis zu 10 Gigabit |

Abbildung 2
On-Demand Preise für Amazon EC2 ¹⁵

Es ist zu beachten, dass Instanzen mit denselben Eigenschaften, aber in verschiedenen Regionen, unterschiedliche Preise haben können.

3.2 Reservierte Instanzen und Saving Plans

Die Zahlungsmodelle Reservierte Instanzen und Saving Plans sind sich sehr ähnlich. Beide kommen mit einer gleichbleibenden Nutzungsverpflichtung, die in €/Stunden gemessen wird. Um die reduzierten Preise zu bekommen, müssen Verträge über ein oder drei Jahre abgeschlossen werden.

Die Abbildung 3 zeigt die möglichen Einsparungen je nach Zahlungsmodell. Die Einsparungen hängen mit der Flexibilität bei der Wahl der Instanzfamilie und der Verfügbarkeitszone zusammen, in die Instanzen übertragen werden können. Je geringer die Flexibilität, desto höher die Einsparungen.

Compute Saving Plans¹⁷ bieten die Flexibilität EC2-Instanzen nach Familie¹⁸, Größe, Verfügbarkeitszone (AZ), Betriebssystem oder Mandant zu wechseln. Diese Option ist bei

¹⁷[12]AWS Saving Plans Pricing

¹⁸[2], WS Certified Solutions Architect - Associate (SAA-C02), Seite 95

| Mögliche Einsparungen laut AWS | | | |
|--------------------------------|---------------------------|--------------------------------|-----------------------------|
| Saving Plans | | Reserved Instances | |
| Compute Saving Plans | EC2-Instance Saving Plans | Convertible Reserved Instances | Standard Reserved Instances |
| bis zu 66% | bis zu 72% | bis zu 54% | bis zu 72% |

Abbildung 3

Mögliche Einsparungen bei Reserved Instances and Saving Plans laut AWS ¹⁶

EC2-Instance Saving nicht möglich und daher bietet diese Alternative eine etwas höher Einsparung.

„Bei Compute Saving Plans können Sie beispielsweise jederzeit von C4- auf M5-Instances wechseln, eine Workload von EU (Irland) nach EU (London) verlagern oder eine Workload von EC2 auf Fargate oder Lambda verschieben. Dabei zahlen Sie automatisch weiterhin den Saving Plans-Preis.“ ¹⁹

Bei den EC2-Instance Saving Plans hingegen muss eine Instance-Familie in einer bestimmten Region ausgewählt werden. Dies reduziert automatisch die Kosten für die ausgewählte Instanz-Familie in der jeweiligen Region, unabhängig von Availability Zone, Größe, Betriebssystem oder Mandant.

EC2 Reserved Instance Marketplace

Sollte sich herausstellen, dass die Kapazität der reservierten Instanzen viel zu wenig oder gar nicht genutzt wird, kann diese Rechenkapazität auf dem RI Marketplace zur Verfügung gestellt werden. Dadurch könnte Teil der Investition wieder hereinzuholen. Dies ist für Standard Reserved Instances möglich. Diese Instanzen werden in Spot-Instanzen umgewandelt damit andere Nutzer sie beantragen können. Eine Servicegebühr sollte in Betracht gezogen werden. Im November 2021 beträgt diese Abgabe 12%²⁰. [Rev]

Möglichkeit der Vorauszahlung

Zusätzlich gibt es bei Saving Plans und reservierten Instanzen die Option im Voraus zu zahlen. Im Gegenzug wird ein niedrigerer Preis angeboten. Amazon bietet drei verschiedene Optionen an. Diese sind teilweise, keine oder vollständige Vorauszahlung²¹. Bei teilweiser Vorauszahlung ist eine Anzahlung von etwa 50% zu leisten.

¹⁹[12]AWS Saving Plans Pricing

²⁰[24]Amazon EC2 Reserved Instance Marketplace

²¹[18] AWS Pricing Calculator

Die Abbildung 4 zeigt den Vergleich zwischen den drei Optionen für Vorauszahlungen. Hier wird deutlich, dass es kaum einen Unterschied zwischen einer teilweise Vorauszahlung und keine Vorauszahlung zu machen gibt. Eine erhebliche Einsparung ergibt sich, wenn man für den gesamten Zeitraum der gebuchten Instanzen im Voraus bezahlt.

| Zahlungsmodell | | EC2 Instance Saving Plans | |
|------------------------|--------------------|---------------------------|--------------------|
| Anzahl der Instanzen | 20 | | |
| Dauer | 36 | Monate | |
| Vorauszahlung | keine | teilweise | vollständig |
| Gesamtkosten pro Monat | \$967.98 | \$519.62 | \$0.00 |
| Vorabkosten gesamt | \$0.00 | \$16,135.92 | \$30,327.12 |
| Gesamtbetrag | \$34,847.28 | \$34,842.24 | \$30,327.12 |
| Prozentuale Einsparung | - | 0.01% | 12.96% |
| Monetäre Einsparung | - | \$5.04 | \$4,515.12 |

Ohne Elastic Block Storage (EBS)

Abbildung 4

Mögliche Einsparungen durch Vorauszahlungen für EC2 Instanzen in Saving Plans
Zahlungsmodell

Eigene Darstellung. Quelle: [18]AWS Pricing Calculator

Die Berechnungen wurden mit dem AWS Pricing Calculator [18] für Instanzen der Familie t4g.xlarge, in der EU (Frankfurt) und für eine Laufzeit von 3 Jahren durchgeführt.

3.3 Spot Instanzen

Wie in Unterkapitel 3.2 genannt bieten EC2 Spot-Instanzen die Möglichkeit aus den ungenutzten EC2-Instanzen anderer Nutzer zu profitieren. Mit einem Preisvorteil von bis zu 90 % gegenüber normalen On-Demand-Instanzen sind Spot-Instanzen ideal für fehlertolerante Anwendungen wie auf Containern ausgeführte Workloads, CI/CD, Bigdata-Anwendungen und ähnliches.

Unterbrechbarkeit

Es ist zu beachten, dass Spot-Instanzen jederzeit unterbrochen werden können. Einer der Gründe ist die Preisüberschreitung der Instanz. Wenn Spot-Instanzen angefordert werden, wird ein Maximalpreis festgelegt. Ist der Preis der Spot-Instanz höher als der eingegebene Maximalpreis, ist die Spot-Instanz für die aktuelle Einstellung nicht mehr

verfügbar. Ein anderes Szenario ist, wenn der Instanz Anbieter die Spot-Instanz erneut anfordert. Falls eine Spot-Instanz unterbrochen wird, benachrichtigt Amazon EC2 zwei Minuten im Voraus. Dieses Ereignis ist verfügbar auf CloudWatch, damit weitere Alarmen eingestellt werden. Diese und andere Funktionalitäten von CloudWatch werden in Kapitel 4 näher erläutert.

Da Spot-Instanzen anfällig für Unterbrechungen sind, ist es nicht empfehlenswert, für Produktionsumgebungen nur Spot-Instanzen zu verwenden.

3.4 Amazon EC2 Fleet[rev]

Instanzen-Flotte oder auf Englisch *fleet of instances*, bieten bei AWS die Möglichkeit mehrere Spot-Instanzen anzufordern, um einen bestimmten Bedarf an Rechenleistung zu decken²². Spot-Instanzen können auch für produktive Umgebungen verwendet werden²³. Darüber hinaus ist es empfehlenswert, Instanzen aus verschiedenen Zahlungsmodellen zu kombinieren, um von den Einsparungen von Spot-Instanzen, Saving Plans und reservierten Instanzen zu profitieren. Die Kombination von Instanzen aus verschiedenen Zahlungsmodellen beseitigt den Nachteil für Produktionsumgebungen, der mit Spot-Instanzen verbunden ist. Das heißt, das Risiko eines vollständigen oder teilweisen Ausfalls der Rechenkapazität durch den Ausfall von Spot-Instanzen.

Folgende Punkte sind für die Nutzung von Spot Fleet Instanzen zu berücksichtigen:

Wahl der Spot-Instanzen[rev]

Die Instanzen, die in der Auswahl für die Instanzen-Flotte berücksichtigt werden, müssen die Anforderungen der Applikation entsprechen. Um die Wahrscheinlichkeit zu erhöhen, dass mehr Spot-Instanzen gefunden werden, ist es empfehlenswert, die Kriterien der Suche zu erweitern. Das erreicht man, indem man Instanzen ähnlicher Typen einschließt. Die Berücksichtigung von Instanzen von Familien mit mehr Leistung als erforderlich ist ebenfalls eine gute Option[25]. Denn, obwohl die Leistung die Anforderungen der Applikation überstiegen werden, wird es einen reduzierten Preis für die Spot-Instanzen bezahlt als bei On-Demand Zahlungsmodell.

²²[27] Amazon Elastic Compute Cloud - Benutzerhandbuch für Linux-Instances, Seite 708

²³[25] Running Web Applications on Amazon EC2 Spot Instances

Maximaler Stundenpreis[rev]

Wie im Unterkapitel 3.3 erwähnt, muss für die Anforderung von Spot-Instanzen ein Maximalpreis festgelegt werden. In diesem Fall ist die Festlegung dieses Maximalpreises auch für die gesamte Instanzen-Flotte eine Option. Es kann erwartet werden, dass die Spot-Preise im Laufe der Zeit stabil bleiben, das heißt keine starken Preisschwankungen. Der aktuelle Preis und der Preisverlauf von Spot-Instanzen können in der AWS-Konsole²⁴ abgefragt werden. Diese Informationen sind zugänglich nur mit einem AWS-Konto.

Festlegung von On-Demand-Anteil[rev]

Wenn alle oder eine große Anzahl von Spot-Instanzen nicht mehr verfügbar sind, muss die benötigte Rechenkapazität von Instanzen anderer Zahlungsmodellen wie On-Demand abgedeckt werden. Die Standardeinstellungen liegen bei 70% On-Demand-Instanzen und 30% Spot-Instanzen[25]. Im Falle von vorhandenen reservierten Instanzen oder Instanzen von Saving Plans werden On-Demand-Instanzen zum entsprechend reduzierten Preis berechnet²⁵.

Auto Scaling Groups[rev]

Auch als EC2-Auto-Scaling-Gruppe(ASG) bezeichnet, ist für die Skalierung der zu startenden Instanzen verantwortlich. Dazu muss vorher eine Startkonfiguration erstellt werden. Dies definiert die Konfiguration der zu startenden Instanzen. In der Startkonfiguration werden unter anderem der Instanztyp, Security-Groups, und Tags festgelegt. Mehr über Auto-Scaling und seine verschiedenen Konfigurationen in Kapitel 5.

Für die Nutzung von EC2-Flotten und Auto Scaling-Gruppen fallen keine zusätzlichen Kosten an. Man muss nur für die durch die EC2-Instanzen verursachten Kosten bezahlen²⁶.

3.5 Anwendungsfall: TrueCar[rev]

Instanzen in Zahlungsmodellen, die zu zeitlichen Verpflichtungen führen, birgt die Gefahr, dass die benötigte Rechenkapazität mittel- bis langfristig falsch eingeschätzt wird. Einerseits kann die reservierte Rechenkapazität zu gering eingeschätzt werden. Als Konsequenz

²⁴[26] AWS EC2 Spot Instanzen-Anfragen und Preisverlauf

²⁵[27] Amazon Elastic Compute Cloud - Benutzerhandbuch für Linux-Instances, Seite 690

²⁶[27] Amazon Elastic Compute Cloud - Benutzerhandbuch für Linux-Instances, Seite 709

wird es größtenteils der Rechnerkapazität mit On-Demand-Instanzen gedeckt, welche in dem Anteil der reservierten Instanzen berücksichtigt werden konnten und mit reduzierten Preisen berechnet. Andererseits, wenn zu viel Rechnerkapazität mit reservierten Instanzen reserviert und diese zu wenig gebraucht wird. Besteht die Möglichkeit, dass es die reine Nutzung von On-Demand-Instanzen eine kostengünstigere Option darstellt.

Im Folgenden wird die Strategie beschrieben, dass TrueCar Inc. verfolgt hat, um in keine der beiden oben genannten Situationen zu geraten. Dank ihrer Optimierungsstrategie konnten sie ihre AWS-Kosten durch die Nutzung reservierter Instanzen um etwa 40% senken²⁷.

Um Einsparungen von 40% zu erreichen, musste das Team von TrueCar zuerst verstehen, wie AWS-Dienste wie reservierte Instanzen, Cost-Explorer, Auto-Scaling-Gruppen und Lambda Funktionen funktionieren. Damit haben sie eines der häufigsten Hindernisse überwunden, mit denen Unternehmen bei der Nutzung von Cloud-Diensten konfrontiert werden und zwar die Mangel an technisches Wissen in Bezug auf Cloud-Dienste²⁸. Nachdem das Team von TrueCar die notwendigen Informationen, insbesondere über die reservierten Instanzen, verstanden haben, haben sie die benötigte Rechenkapazität ermittelt. In dem Artikel wurde nicht erläutert, wie die von TrueCar benötigte Rechnerkapazität berechnet wurde. Diese Informationen werden jedoch von Cost-Explorer bereitgestellt. Cost-Explorer bietet die Möglichkeit, die Nutzung der AWS-Services für die letzten 12 Monate anzuzeigen. Cost-Explorer wird in Unterkapitel 4.2 ausführlicher behandelt.

Die Kosten der Instanzen in On-Demand wurden mit dem von reservierten Instanzen gegenübergestellt, um den Break-Even-Point dazwischen zu finden. Der Break-Even-Point bedeutet in diesem Fall, der Punkt, wo die Preise der reservierten Instanzen und die On-Demand Instanzen gleich sind. Nach diesem Punkt wird der monatliche Preis für die reservierten Instanzen sinken, bis die reservierte Kapazität verbraucht wird oder der Zeitraum für die reservierten Instanzen endet.

Wie in der Grafik der Abbildung 5 dargestellt wird liegt der Break-Even-Point zwischen dem Monat acht und neun. Im Fall, dass da Verbrauch der Instanzen vor dem Monat auch endet, würde, wäre es nicht empfehlenswert Instanzen zu reservieren, sondern mit On-Demand Instanzen zu arbeiten. Die Berechnung wurde gemacht für den Zeitraum

²⁷[37]How TrueCar Saves 40% on AWS with EC2 Reserved Instances

²⁸[1], Accenture Dienstleistungen GmbH. Hohe Erwartungen an die Cloud: Hürden meistern, Mehrwert maximieren, Seite 11

von 1 Jahr durchgeführt. [RECHNEN UND ERKLÄREN]. In dem Prozess wurden die

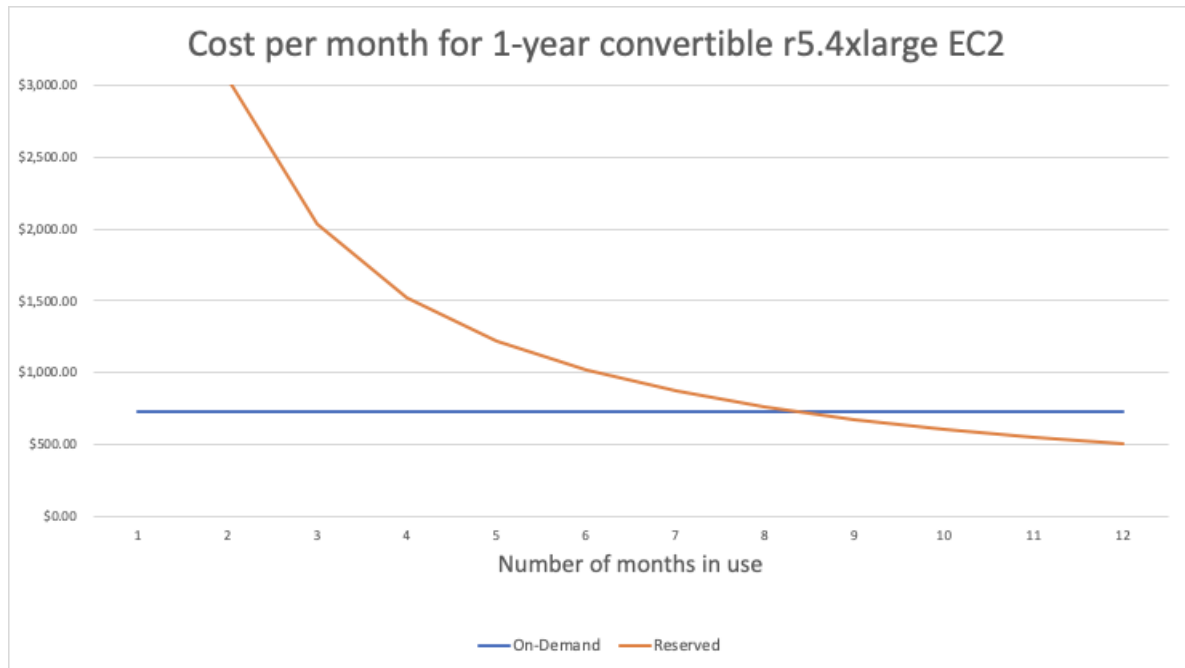


Abbildung 5
Monatliche Kosten für eine On-Demand-Instanz
im Vergleich zu einer reservierten Instanz.
Quelle: [37]

Buchhaltungs- und Finanzabteilungen involviert[WICHTIG WEIL], um die Preisvorteile zu besprechen. Nach der Buchung der reservierten Instanzen wurde deren Nutzung mit Cost-Explorer überwacht.

Mit Cost-Explorer wurden die folgenden 2 Metriken überwacht:

RI-Coverage, die anzeigt, wie viel der On-Demand-Instanzen durch reservierten Instanzen abgedeckt wird. Ziel ist hierbei das RI-Coverage der reservierten Instanzen so nahe wie möglich an 100% zu halten.

RI-Utilization, welche zeigt, wie viel Prozent der reservierten Instanzen verbraucht wurden. Es wird versucht die RI-Utilization nicht zu niedrig zu halten.

Um diese Metriken im Blick zu behalten und nicht jeden Tag den Cost-Explorer aufrufen zu müssen, wurde eine Benachrichtigung an Slack eingerichtet. Dies war über die Cost-Explorer API und eine Lambda-Funktion möglich.

TrueCar, Inc. ist eine Preis- und Informations-Website für Neu- und Gebrauchtwagenkäufer mit Sitz in Santa Monica, Kalifornien²⁹.

Vergleich der Zahlungsmodelle[Rev]

Die folgende Tabelle fasst die Eigenschaften der Zahlungsmodellen für On-Demand-, reservierte, von Saving Plans und Spot-Instanzen zusammen und listet typische Applikationen je nach Zahlungsmodell auf. [Abb. VOLLSTÄNDIG?AKTUELL?]

| Vergleich der Zahlungsmodelle | | |
|---|---|---|
| Eigenschaften | | |
| Nutzungsabhängige Zahlung: On-Demand | Optionen mit Verpflichtung: Reserved Instances and Saving Plans | Überschüssige Kapazität: Spot-Instances |
| Erster Test oder erste Entwicklung | Verträge über 1 bis 3 Jahre | Unterbrechbare Instanzen |
| Keine langfristigen Verpflichtungen | Preisverpflichtung | Die billigste und riskanteste Option |
| Keine Vorabzahlungen | | |
| | | |
| Geeignete und übliche Anwendungen | | |
| Allgemeine Anwendungen | Applikationen mit stabiler Arbeitsbelastung | Bigdata-Applikationen |
| Experimente und Tests | | Containern ausgeführte Workloads |
| Nicht unterbrechbare Applikationen | | Fehlertolerante Applikationen |
| Applikationen mit unvorhersehbaren Arbeitsbelastungen | | Batch-Workloads |
| | | |

Abbildung 6
Vergleich der Zahlungsmodelle nach Eigenschaft und Anwendungsfall
Eigene Darstellung. Quelle: [4, 8, 12, 20, 41]
[38]Plusserver: Kostenoptimierung in AWS Seite 9

²⁹Die Quelle dieser Informationen ist ein Artikel, der auf <https://www.medium.com> veröffentlicht wurde. Dass der Artikel von TrueCar stammt, wird durch die Tatsache bestätigt, dass deren Website <https://www.truecar.com/who-we-are/> zu dem hier erwähnten Artikel führt.

Fazit[Rev]

In diesem Kapitel wurden die verschiedenen Zahlungsmodelle für EC2-Instanzen untersucht. Es wurden Hinweise für die Auswahl des richtigen Zahlungsmodells in verschiedenen Szenarien gegeben. Dies wurde erklärt, um die Preisvorteile von den Zahlungsmodellen zu nutzen. Beginnend mit dem On-Demand-Zahlungsmodell, gefolgt von Reserved Instanzen und Saving Plans. In dieser Reihenfolge sinkt der Preis und mit ihm steigt die Verpflichtung, sich langfristig zu binden. Schließlich mit Spot-Instanzen, die die niedrigsten Preise bieten, aber keine volle Verfügbarkeit sicherstellen.

Im nächsten Kapitel wird CloudWatch[UND...] vorgestellt, mit dem überprüft werden kann, ob das ausgewählte Zahlungsmodell tatsächlich das Richtige für den betreffenden Anwendungsfall ist. Für das On-Demand-Zahlungsmodell gibt es keine Kostenreduzierung, aber es gibt Maßnahmen, um die Nutzung von Instanzen zu reduzieren. Auf diese Maßnahmen wird im Kapitel 5 näher eingegangen.+Cost-Explorer+Trusted Advisor.

4 Kostenüberwachung

In diesem Kapitel werden Werkzeuge vorgestellt, mit denen Budgets mit Alarmen erstellt werden, diese informieren die Nutzer, wenn ein bestimmter Prozentsatz des festgelegten Budgets überschritten wurde. Die Erstellung von Budgets trägt zu einer besseren Planung-/Prognose- und Kostenkontrolle. [WEIL, Cost Controlling] Die Einstellung von Alarmen für relevante Ereignisse wie im Fall einer Budgetüberschreitung oder dem Start einer Instanz[IST GUT WEIL/TRÄG ZU...BEI].

Darüber hinaus ist es mit Werkzeugen wie CloudWatch möglich, die Abschaltung bestimmter Dienste zu automatisieren, wenn eine Budgetschwelle überschritten wurde. Diese Maßnahmen werden in dem nächsten Kapitel genauer behandelt. [TRUSTED ADVISOR fehlt noch in der Einleitung] [SOLLTE AB HIER EINE NEUE UNTERKAPITEL ANFANGEN? zb. TAGS ZUR TRENNUNG DER DIENSTEN] Durch die Verwendung von Tags ist es möglich, Diensten nach Kriterien wie Region, Umgebung, Projekt, Art des Dienstes usw. zu visualisieren, dies ermöglicht, Kosten auf den von der Organisation festgelegten Ebenen zu verfolgen.

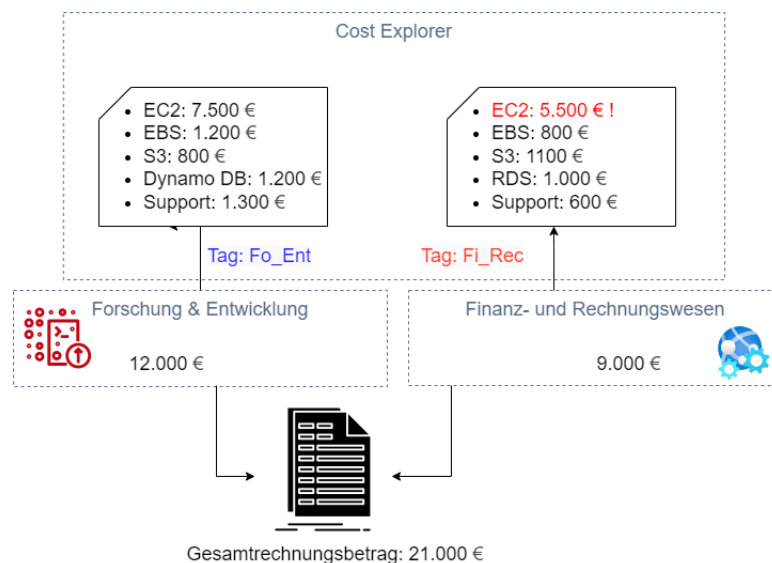


Abbildung 7
Trennung der Kosten durch Tags.

Die Angaben dienen nur als Beispiel und entsprechen keiner realen IT-Infrastruktur.

Es könnte zum Beispiel ein Szenario entstehen, in dem eine Abteilung innerhalb einer Organisation mehr Kosten verursacht als Andere. Dies ist nur durch den Anstieg der von Amazon generierten Rechnung, aber um den Grund für diesen Anstieg genauer zu verstehen, muss ihre Ursache untersucht werden. Werkzeuge wie Cost-Explorer machen

diese Art von Analyse möglich.

4.1 AWS CloudWatch

Amazon CloudWatch ermöglicht die Überwachung der Leistung von Diensten, auch bei Diensten, die über verschiedene Regionen verteilt sind. CloudWatch sammelt operative Daten für die Verlaufsanalysen und die Entscheidungsfindung in Bezug auf Optimierung und Fehlerbehebung. CloudWatch beschränkt sich nicht nur darauf, Daten aus der AWS-Umgebung zu empfangen. Externe Metriken, die mit CloudWatch kompatibel sind, können für eine einheitliche Analyse aggregiert werden. Eine der Metriken, die mit Amazon CloudWatch überwacht werden kann, ist die CPU-Auslastung von EC2-Instanzen. Basierend auf einem Prozentsatz der CPU-Auslastung können Benachrichtigungen und Aktionen konfiguriert werden. Eine dieser Aktionen ist die automatische Einrichtung neuer Instanzen zur Deckung des Kapazitätsbedarfs³⁰. Diese Art von Aktionen werden im Kapitel 5 tiefer behandelt.

Im Folgenden werden die grundlegenden Bereiche und Begriffe von CloudWatch erläutert und wie sie zur Überwachung von Informationen über AWS-Dienste verwendet werden.

Metriken

Eine Metrik stellt eine Reihe von Daten über die Leistung eines Dienstes in zeitlicher Reihenfolge dar. Standardmäßig werden viele kostenlose Metriken an CloudWatch übermittelt. Zum Beispiel kann der Durchschnitt von einer bestimmten API pro Stunde untersucht werden. Für eine detailliertere Überwachung ist es möglich, benutzerdefinierte Metriken zu konfigurieren, die eine Auflösung von bis zu eine Sekunde zulassen. Ein praktisches Beispiel für benutzerdefinierte Metriken ist die Messung der Ladezeit einer Website. [ANDERES BEISPIEL?]

Ereignisse

Bei CloudWatch wird eine Änderung bei einer Ressource in der AWS-Umgebung Ereignis genannt. AWS-Ressourcen können Ereignisse erzeugen, wenn sich ihr Status ändert. Beispielsweise, wird ein Ereignis erzeugt, wenn Amazon EC2 Auto Scaling, Instanzen gestartet oder beendet werden³¹ oder wenn eine bestimmte Menge an Speicherplatz in einem Bucket erreicht wurde. Ein Bucket ist ein Behälter, in dem Objekte bei Amazon

³⁰[2], AWS Certified Solutions Architect - Associate (SAA-C02), Seite 185

³¹[14], AWS Cloud Watch Events: User GuideSeite 1

S3 gespeichert werden³². Beispiele für Objekte sind Dateien wie Bilder und Videos.

Regel

Eine Regel ordnet eintreffende Ereignisse zu und leitet diese zur Verarbeitung an Ziele weiter. Eine einzelne Regel kann an mehrere Ziele weiterleiten, die alle parallel verarbeitet werden³³.

Ziele

Ziele oder Targets sind Ressourcen, die aufgerufen werden, wenn eine Regel ausgelöst wird. EC2 instances, AWS Lambda functions und Amazon SNS sind unter anderem mögliche Ziele. Die Ziele einer Regel müssen sich in derselben Region wie die Regel befinden³⁴.

Benachrichtigungen

Benachrichtigt zu werden ist wichtig, um relevante Ereignisse nicht zu verpassen und rechtzeitig Maßnahmen zu ergreifen. Mit CloudWatch können Alarme eingerichtet werden, die durch Metriken wie die CPU-Auslastung und auch Gebühren[PASSENDEN WORT?] auf AWS-Rechnungen ausgelöst werden. Benachrichtigungen können durch Amazon SNS³⁵ oder zu einer E-Mail-Adresse geschickt werden.

Visualisierung von Metriken

Mit Cloud-Watch Dashboards können relevante Metriken grafisch dargestellt werden. Durch die Dashboards können auch Benachrichtigungen erstellt werden. Für die Einrichtung der Benachrichtigungen ist kein technisches Wissen nötig³⁶. Die in den Dashboards enthaltenen Informationen sind nicht nur für ihre Autoren von Relevanz. Weitere Personen innerhalb oder außerhalb der Organisation können Zugriff auf Dashboards mit nützlichen Informationen bekommen, um Prozesse zu beschleunigen[VORTEILE DES OPTIMIERUNGSPROZESS VERWEISEN] oder Probleme schneller zu beheben[USE CASE].

Dies ermöglicht einen schnelleren Kommunikationsfluss in Echtzeit[Fach Informationsmanagement: (Schwachstellenanalyse) IuM Technik/Check Book KCrmar]. Die Zugriffsverwaltung für geteilte Dashboards wird über AWS Identity and Access Management abgewickelt³⁷[UND?].

³²[19]Amazon Simple Storage Service User Guide, Seite 4

³³[14], AWS Cloud Watch : User Guide, Seite 2

³⁴[14], Seite 2

³⁵Amazon SNS ist ein Dienst, der die Benachrichtigung an Personen und an Applikationen ermöglicht.
<https://aws.amazon.com/de/sns/>

³⁶[15],AWS Cloud Watch : User Guide, Seite 28

³⁷[15], Seite 18 und 39

Fakturierungsalarme mit CloudWatch

AWS CloudWatch empfängt Abrechnungsmetriken von alle Ressourcen. Auf der Grundlage dieser Metriken ist es daher möglich, Regeln zu erstellen, die bei Überschreitung des geplanten Budgets einen Alarm auslösen. [BEZUG AUF PROJEKTMANAGEMENT/-Kostenkontrolle?]

Benachrichtigung bei Hoch- und Runterfahren von EC2-Instanzen

Obwohl Auto-Scaling dafür sorgt, die Rechenkapazität dynamisch anzupassen, ist es von größter Wichtigkeit, über Änderungen in der Infrastruktur informiert zu sein, ohne die Dashboards manuell überprüfen zu müssen. [WEIL]

4.2 AWS Cost-Explorer

Mit Cost-Explorer können Kosten der letzten 12 Monate und eine Schätzung der Kosten des laufenden Monats visualisiert werden. Darüber hinaus wird eine Kostenprognose für die nächsten Monate erstellt. Die Prognose basieren auf die Kosten der vergangenen Monaten. Die Nutzung des Cost-Explorers ist kostenlos, nur API-Aufrufe sind kostenpflichtig³⁸.

Amazon analysiert die bisherige Nutzung der Instanzen und gibt Empfehlungen zur Kostensenkung durch den Wechsel von EC2-Instanzen in On-Demand Zahlungsmodell zu reservierten Instanzen. Diese ignorieren Kapazität, die bereits von anderen reservierten Instanzen abgedeckt wurden. Es besteht die Möglichkeit, einen Bericht über die Ressourcennutzung der letzten 12 Monate einzusehen und zu prüfen, ob die Empfehlungen des Cost-Explorers angemessen sind.

Budgetplanung

Die Budgetplanung ist eine Methode der Kostenkontrolle, die beim Start eines neuen Projekts eingesetzt wird³⁹. Der Bericht über die in den letzten 12 Monaten entstandenen

³⁸[23]AWS Cost Management Pricing

³⁹[31]Cost Control Methods: Definitions and Examples

Kosten zusammen mit der Prognose der Kosten der kommenden drei Monaten tragen zu einer guten Budgetplanung bei. Durch die Möglichkeit, die in den letzten Monaten angefallenen Kosten nach Ressourcen, Projekt oder Abteilung zu trennen, ist es möglich, operative Budgetplanungen aus vergangenen Projekten mit Genauigkeit zu erstellen.

„Bei der operativen Planung wird von einem Zeithorizont von einem Jahr ausgegangen. Hier liegt der Fokus darauf, Ressourcen konkret zuzuweisen und detaillierter zu planen. Welche Mittel werden wofür verwendet und welche kurz- und mittelfristigen Ziele sollen durch diesen Mitteleinsatz erreicht werden“[30].

Cost-Explorer liefert Informationen zur Rechtfertigung von Ausgaben aus im Voraus festgelegten Budgets, hilft bei der Planung künftiger Budgets und unterstützt die []

4.3 AWS Trusted Advisor

AWS Trusted Advisor ist ein Werkzeug, das entwickelt wurde, um Kosten zu senken, um Systemverfügbarkeit und -leistung zu verbessern und um Sicherheit zu erhöhen. Es analysiert die Nutzung des AWS-Kontos und gibt Best-Practice-Empfehlungen. Es werden die Kategorien Leistungsgrenzen und Kostenoptimierung insbesondere betrachtet, da diese am relevantesten für die vorliegende Arbeit sind. Es ist zu berücksichtigen, dass nur limitierte Sicherheitsprüfungen (6 Prüfungen November 2021) für Konten in den Plänen Developer und Basic Support kostenlos sind. Prüfungen für die Kategorie Leistungsgrenzen sind kostenlos. Detaillierte Informationen und Empfehlungen von der Kategorie Kostenoptimierung, Performance und Fehlertoleranz sind nur zugänglich, wenn ein Business- oder Enterprise-Konto vorliegt⁴⁰.



Abbildung 8
AWS Trusted Advisor Kategorien[21]

⁴⁰<https://aws.amazon.com/de/premiumsupport/technology/trusted-advisor/>

Die Abbildung 8 zeigt die fünf Kategorien von Trusted Advisor mit jeweils 3 Arten von Indikatoren. Die Indikatoren zeigen an, welche Prüfungen durchgeführt wurden. Grün bedeutet, dass keine Fehler oder zu prüfenden Empfehlungen vorhanden sind. Warnungen werden durch orangefarbene Indikatoren und Fehler durch rote Indikatoren angezeigt. Diese Empfehlungen sind eine Zusammenfassung auf hohem Niveau. Sie sind ein Startpunkt für die Untersuchung von Ressourcen mit Hilfe anderer Werkzeuge wie CloudWatch oder Cost-Explorer.

Kategorie Kostenoptimierung

Die Empfehlungen zur Kostenoptimierung konzentrieren sich auf Möglichkeiten zur Kostensenkung, indem ungenutzte Ressourcen hervorgehoben werden. Sollten EC2-Instanzen mit geringer Auslastung gefunden werden, wird es diese bei Trusted Advisor signalisiert. Denn diese Instanzen verbrauchen Ressourcen und können terminiert oder pausiert werden. Auch nicht zugewiesene Elastic IP-Adressen erzeugen Kosten. Diese können gegebenenfalls von Trusted Advisor gefunden werden. [BEISPIELE]

Kategorie Leistungsgrenzen

In dieser Kategorie werden Empfehlungen zur Vermeidung von Grenzwertüberschreitungen hervorgehoben. Es wird zum Beispiel nach einer Nutzung gesucht, die mehr als 80 % des Leistungsgrenzwerten für wichtige Dienste beträgt. Einige Beispiele sind Amazon EC2, Auto Scaling, Elastic Block Store, Simple Email Service und AWS CloudFormation.

Sich dieser Grenzen bewusst zu sein, gibt die Möglichkeit, rechtzeitig zu handeln und es trägt zu Kostenüberwachung bei. [GLEICHE GRENZEN WIE BEI CloudWatch?] Bei der Erwägung von Trusted-Advisor ist zu überlegen, ob es kosteneffizient ist, für Pläne zu zahlen, die den Zugang zu allen Empfehlungen des Trusted Advisors ermöglichen. Das übergeordnete Ziel dieser Arbeit ist es, die Entstehung der Kosten auf eine praktikable Weise zu verstehen (Kostenüberwachung). Dies, um Optimierungsmaßnahmen zu ermöglichen.

Es wäre nicht sinnvoll, Kosten für Plänen wie Geschäfts- oder Enterprise Support zu übernehmen, wenn diese die möglichen Einsparungen übersteigen. Die Vorteile von Geschäfts- oder Enterprise Support-Plänen beschränken sich nicht auf Kosteneinsparungen und Kostenbegrenzung, sondern tragen auch zur Sicherheit und Leistung bei. Jedes Unternehmen

muss selbst entscheiden, ob es diese Informationen benötigt. [ZEIGE BEISPIELE FueR Empfehlungen]

4.4 Überwachungswerkzeuge gemäß ihrer Verwendung

Die Abbildung 9 fasst die Überwachungswerkzeuge zusammen und listet die Verwendung der einzelnen Werkzeuge auf. [NOCH NICHT VOLLSTÄNDIG]

| Überwachungswerkzeuge gemäß ihrer Verwendung | | | |
|--|-------------|---------------|-----------------|
| | Cloud-Watch | Cost-Explorer | Trusted-Advisor |
| Visualisierung der CPU utilization | x | | |
| Analyse von Kosten nach Tags, Monat... | | x | |
| Benachrichtigung/Alarmen von Events | x | | |
| Empfehlungen bezüglich RIs | | x | x? |
| Um Ressourcen nach Tag zu | | x | |
| Prognose für kommende Kosten | | | |
| | | | |
| | | | |

Abbildung 9
Überwachungswerkzeuge gemäß ihrer Verwendung
Eigene Darstellung[13, 21, 22].

Handlungsempfehlungen

[SIND SIE HIER RICHTIG PLAZIERT? SOLLTEN LIEBER IN FAZIT SEIN?;NOCH ZU VERVOLSTÄNDIGEN]

Handlungsempfehlung 1:

Es kann in Erwägung gezogen werden, für einen begrenzten Zeitraum von 3 Monaten einen Support-Plan zu bezahlen, um aus den gegebenen Empfehlungen zu lernen. Oder Business-Plan alle 6 Monate für 1 Monat zu aktivieren.

Handlungsempfehlung 2:

Ein Berater für eine Prüfung und Optimierung der Ressourcen kann in Deutschland zwischen x und N-EUR kosten. Dies ist eine Alternative zu den Plänen des Trusted-Advisor. Ein Berater, der alle 5 Kategorien abdeckt, könnte [BETRAG] kosten.

Fazit

In diesem Kapitel wurde gezeigt, dass es mit CloudWatch möglich ist, Alarmer auf Basis von Ereignissen einzurichten, die mit Amazon SNS oder externen E-Mail-Adressen kommunizieren. Im nächsten Kapitel wird CloudWatch erneut behandelt. Diesmal nicht als Überwachungswerkzeug, sondern als Optimierungswerkzeug zur Erstellung von automatisierten Aktionen/Reaktionen. Dazu war es notwendig, die Rolle der von CloudWatch gesammelten Metriken zu verstehen, die die Grundlage für die Verwaltung von Aktionen wie Auto-Scaling-Gruppen bilden. Aus dem Blickwinkel des Kostenmanagements wurde gezeigt, dass mit Cost-Explorer eine Analyse von Kosten der letzten 12 Monate, eine Einschätzung der Kosten im aktuellen Monat und eine Prognose für die nächsten Monate möglich ist. Diese Informationen dient unter anderem zur Erstellung einer operativen Budgetplanung mit genaueren Daten, da Kosten nach Tags und anderen Filtern getrennt werden können. Darüber hinaus wurde Trusted Advisor vorgestellt, der konkrete Optimierungsempfehlungen gibt und warnt über Leistungsgrenzen. Dies kann mit erheblichen Kosten verbunden sein und ist daher nicht für alle Arten von Unternehmen unmittelbar attraktiv. Obwohl sich nicht alle Unternehmen die Prüfungen von Trusted Advisor leisten können, sollten die kostenlosen Empfehlungen im Überwachungs- und Optimierungsplan berücksichtigt werden. [WAS KOMMT IN NÄCHTEN KAP.?)

5 Optimierungsmassnahmen

Die mit den Überwachungswerkzeuge gesammelte Informationen, bilden die Grundlage für die Optimierungsmassnahmen.[KONKRETER WERDEN] In diesem Kapitel werden die mit Hilfe der Werkzeuge gewonnenen Informationen genutzt[INFO X FÜR WERKZEUG 1...], um über die am besten geeigneten Optimierungsmassnahmen zu entscheiden.

5.1 EC2 Automatische Skalierung

Auto Scaling ist es hilfreich, um die richtige Anzahl von EC2 Instanzen zur Verfügung zu haben[NÖ+horizontaal/vertikal Skalierung], um die Anwendungslast dynamisch abzudecken.

Die Abbildung 10 zeigt das wechselnde Verhalten einer Beispielanwendung, die vor allem unter der Woche Ressourcen verbraucht. Am Wochenende sinkt die Nachfrage nach Rechenkapazität auf weniger als 25 % und lässt den Rest der Kapazität ungenutzt.

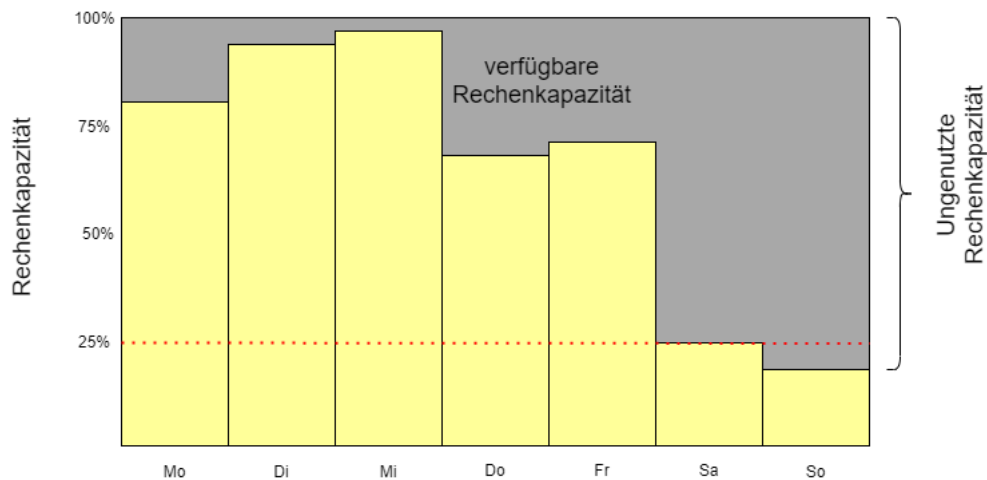


Abbildung 10

Ungenutzte Rechenkapazität ohne automatische Skalierung.

Quelle: Eigene Darstellung mit [fiktiven?] Angaben.

Die gelben Säulen stellen die tägliche genutzte Rechenkapazität dar. Die graue Zone entspricht ungenutzte Rechenkapazität und beträgt etwa ein Drittel der wöchentlichen Rechenkapazität.

Elastic Load Balancing

□

5.1.1 Zeitgesteuerte Skalierung

Nicht produktive Umgebungen

In einem On-Premise-System macht es, wenn überhaupt, einen kleinen Unterschied bei den Kosten, dass Instanzen die ganze Zeit aktiv bleiben. Im Gegensatz dazu ist es bei On-Demand-Zahlungsmodelle sinnvoll Zeiträume zu definieren, in denen Instanzen abgeschaltet werden können, um den Verbrauch an Ressourcen zu reduzieren. Bei Systemen, die nur tagsüber und unter der Woche in Betrieb sein müssen, kann dies eine Einsparung von zu 67% bedeuten. Wenn zum Beispiel Test- und Beta-Umgebungen von Montag bis Freitag von 7 bis 20 Uhr laufen würden.

| Zeitgesteuerte Skalierung von EC2-Instanzen | | |
|--|----------------------------------|------|
| | 7:00-20:00 Uhr Montag-Freitag | 24/7 |
| Stunden inaktiv täglich | 11 | 0 |
| Stunden aktiv täglich | 13 | 24 |
| Tagen in der Woche | 5 | 7 |
| Stunden in der Woche | 55 | 168 |
| Stunden monatlich | 239 | 730 |
| Einsparung/Differenz % | 67.26% | |

| | | |
|------------------------------------|---------|---------|
| Stundensatz | €0.1536 | |
| Anzahl Instanzen | 2 | |
| On-Demand Kosten pro Monat* | €73.42 | €224.26 |

Abbildung 11

Berechnung für ein nicht-produktive Umgebung mit Zeitgesteuerte Skalierung.
Quelle: Eigene Darstellung.

Der Stundensatz wurde am 23.11.2021 mit dem AWS Pricing Calculator[18] ermittelt für Linux Instanzen in Frankfurt mit 4vCPUs, 16 GB Arbeitsspeicher und Instanz-Familie t4g.xlarge in On-Demand-Zahlungsmodell.

Produktive Umgebungen

Wenn der Zeitpunkt einer hohen Nachfrage bekannt ist, kann eine Erhöhung der Rechnerkapazität geplant werden, um Überlastungen zu vermeiden. Beispiele für solche Zeiträume

sind Cyber-Monday und Black Friday⁴¹. [WEITERE ERKLÄRUNG]

5.1.2 Dynamisches Auto Scaling

Es kann jedoch zu schnelle und kontinuierliche Änderungen im Verhalten von Applikationen geben, häufig innerhalb von wenige Minuten. Bei solche Szenarien ist sinnvoller, Metriken zur automatischen Anpassung der Skalierung der Rechenkapazität festzulegen. Beispiele für eine veränderte Nutzung von Applikationen finden sich bei Tinder und OkCupid, zwei der größten Dating-Applikationen in den vereinigten staaten.

Die Abbildung 12 zeigt die Nutzungsspitzen bei den genannten Applikationen. Dieses wechselnde Verhalten wirkt sich unmittelbar auf die zu verschiedenen Tageszeiten benötigte Rechenkapazität aus und macht eine dynamische Skalierung der Rechenkapazität erforderlich, wenn das Ziel darin besteht, die Verschwendung von Ressourcen zu vermeiden oder zu verringern. Die für die automatische Skalierung erforderlichen Metriken wurden bereits im Kapitel Überwachungswerkzeuge erwähnt. Eine der Metriken, die von Cloudexperten [TELEFONAT mit Deepak von der Einheit CIS[Cloud Capgemini]]benutzt wird, ist die gesamte CPU-Auslastung. Um die CPU-Auslastung als Metrik zu verwenden, werden mindestens zwei Schwellenwerte definiert. Eine für die Erhöhung von Rechenkapazität, Scale-Out genannt und eine für das Verringern von Rechenkapazität bezeichnet als Scale-In.

5.1.3 Manual Scaling

Für die Konfiguration einer Auto-Scaling-Gruppe werden die minimale, maximale und gewünschte Anzahl von Instanzen definiert. Wenn aufgrund von Bedingungen, die in der Konfiguration einer Auto-Scaling-Gruppe nicht berücksichtigt wurden mehr Rechenkapazität benötigt wird, ist es möglich, die Rechenkapazität manuell zu steuern. Dies geschieht, ohne dass die aktiven Instanzen unterbrochen werden.

5.1.4 Predictive Scaling

Voraussagende Skalierung oder Predictive Scaling auf Englisch, nutzt maschinelles Lernen, um den Kapazitätsbedarf auf der Grundlage historischer Daten von CloudWatch vorherzusagen. Mit Hilfe der Predictive Scaling kann es die Kapazität vor der erwarteten Auslastung bereitstellen, im Gegensatz zur dynamischen Skalierung, die reaktiv ist. Für Instanzen, die viel Zeit für die Initialisierung benötigen kann die Zeit zwischen dem Beginn

⁴¹[47]Wie viel planen Sie am Black Friday / Cyber Monday auszugeben?

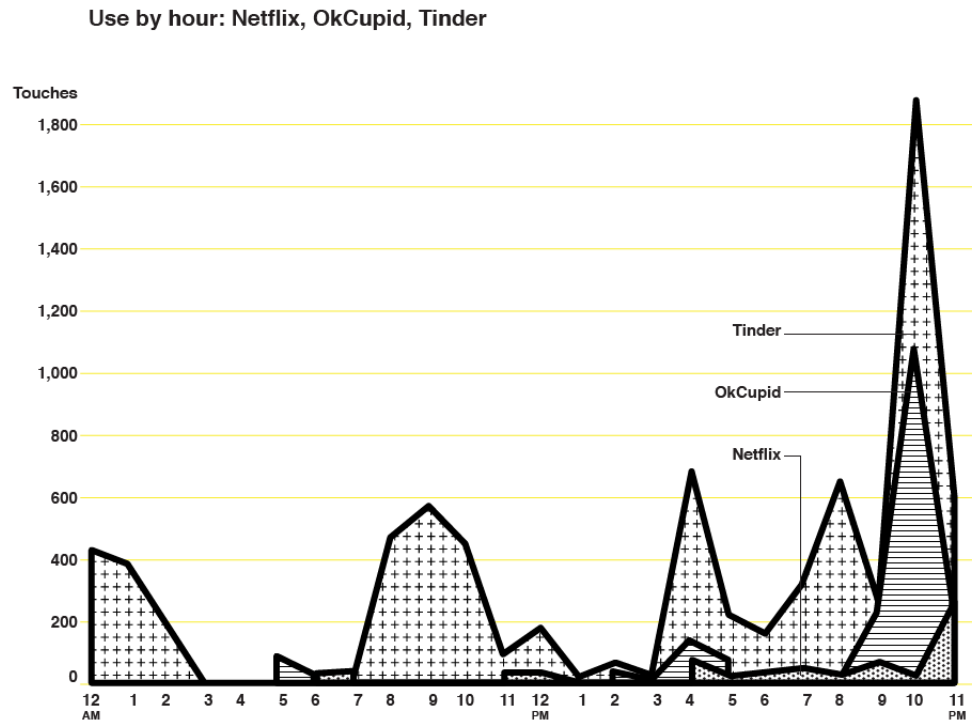


Abbildung 12

Nutzung von Tinder, OkCupid und Netflix pro Stunde. ⁴²

Mit Touches sind die Anzahl der Klicks, Swipes oder einfachen Interaktionen mit der Applikation gemeint.

des Nachfrageanstiegs und der Initialisierung der Instanz vermieden oder verkürzt werden. [(EIGENES) DIAGRAMM] Anders als Zeitgesteuerte Skalierung ist es nicht notwendig, die Verhaltensmuster der Anwendungen zu analysieren.

5.2 S3 Optimierung

5.2.1 Die richtige Speicherklassen wählen

Um die Speicherkosten zu optimieren, ist es daher notwendig, die richtige Speicherklassen für die jeweilige Applikation wählen. Um die richtige Wahl zu treffen, müssen die Anforderungen der Applikation verstanden werden. Klinische Patientendaten und eine Instagram-Story unterscheiden sich in der Zugriffshäufigkeit auf diese Daten und in der Länge der Aufbewahrungszeit.

Amazon bietet verschiedene Speicherklassen an, die sich im Preis und in der Häufigkeit des Zugriffs auf die Objekte unterscheiden. Objekte sind in Behältern enthalten, die

Buckets genannt werden. [ABB Speicherklasse ->Eigenschaften der Anwendung?] Wenn Daten über einen längeren Zeitraum gespeichert werden müssen, weil die Anforderungen der Applikation dies vorschreiben oder für den Fall dass, per Gesetz auf die Informationen in der Zukunft zugegriffen werden muss. [UMFORMULIEREN] Zusätzlich, wenn auf die Daten nicht häufig zugegriffen wird, sind Glacier und Glacier Deep Archive passende Speicherklassen. Die Entscheidung ist jedoch nicht immer so einfach und die Umstände können sich schnell ändern. Hinzu kommt, dass nicht alle Daten in einer Applikation immer die gleichen Zugriffsmuster haben. Für solche Fälle ist es möglich, Regeln zu definieren, die Dateien zwischen verschiedenen Speicherklassen abhängig von ihrem Alter übertragen.

5.2.2 Lebenszyklus-Konfiguration

[MASSNAHME für...]Eine S3-Lebenszykluskonfiguration oder lifecycle policy beschreibt in einer XML-Datei Regeln und Aktionen für die Manipulation von Objekten. Aktionen wie das Verschieben von Objekten verursachen Kosten. Ein Beispiel von diesen Kosten und mögliche Einsparungen werden in Abbildung 13 für die Berechnung der Speicherkosten verwendet.

Um konkretere Regeln zu definieren, ist es möglich Tags zu verwenden und somit eine Unterscheidung zwischen Objekten mit verschiedenen Tags zu treffen. Es ist zum Beispiel möglich, alle Objekte mit dem Tag: Dev nach 45 Tagen nach Standard Infrequent Access und nach 120 Tagen nach S3 Glacier zu verschieben.

```
<LifecycleConfiguration>
  <Rule>
    <ID>example-id</ID>

    <Filter>
      <Tag>
        <Key>key</Key>
        <Value>Dev</Value>
      </Tag>
    </Filter>

    <Status>Enabled</Status>
    <Transition>
      <Days>45</Days>
      <StorageClass>STANDARD_IA</StorageClass>
```

```
</Transition>
<Transition>
  <Days>120</Days>
  <StorageClass>GLACIER</StorageClass>
</Transition>
<Expiration>
  <Days>365</Days>
</Expiration>
</Rule>
</LifecycleConfiguration>
```

Angepasster Code auf Basis der Beispiele auf Seite 701 in
Amazon Simple Storage Service - User Guide,

⁴³ Zur Veranschaulichung (der gezeigten Informationen[OHNE ODER DAMIT?]) wird davon ausgegangen, dass ein Sicherheitsunternehmen, das Sicherheitsvideos speichern muss, im Durchschnitt 120 TB an Videos speichern muss. Viele von ihnen werden mindestens 5 Jahre lang aufbewahrt, falls sie vor Gericht als Beweismittel dienen. Ungefähr 50% der Videos werden mindestens einmal im Monat überprüft und müssen laut Gesetz sofort zugänglich sein. Die Software des Unternehmens speichert die Videos in S3-Buckets und hat eine durchschnittliche Größe von 3,4 GB.

Im Folgenden werden die Speicherkosten für ein Szenario berechnet, bei dem nur S3 Standard verwendet wird. Als nächstes wird die Kombination von S3 Standard Infrequent Access, S3 Glacier und S3 Standard für ein Szenario betrachtet, in dem die Dateien je nach Alter verschoben werden. Im letzten Szenario müssen die Kosten für den Übergang[RICHTIGES WORT?] zwischen Speicherklassen berücksichtigt werden.

Zur Vereinfachung der Berechnung wird angenommen, dass 20% der Dateien in S3 Standard Infrequent Access und 30% in S3 Glacier gespeichert werden.

Bei der Berechnung wurden die Kosten für das Verschieben von Dateien zwischen Speicherklassen berücksichtigt. Anhand der Berechnungen lässt sich erkennen, dass ein Einsparungspotenzial von rund 1.000 Dollar pro Monat besteht, indem die notwendigen Regeln aufgestellt werden, um einen Teil der Dateien in anderen Speicherklassen zu verschieben, welche niedrigere Preise bieten.

⁴³[19], Seite 701

| | |
|------------------------------|---------------------------|
| Durchschnittliche Dateigröße | 3.4 GB |
| Anzahl der Dateien | 36,141 Überwachungsvideos |
| Gesamtspeicher | 122880 GB |
| | 120 TB |

| Ausschließlich S3-Standard verwenden | | |
|--|-----------------------------|------------------------------|
| | S3 Standard (erste 51200GB) | S3 Standard (Nächste 450 TB) |
| Speicherplatz in GB | 51200 | 71680 |
| Preis pro GB | \$0.0245 | \$0.0235 |
| Speicherverteilung | 42% | 58% |
| Anzahl der Dateien | 15059 | 21082 |
| Übertragungsgebühr (pro 1.000 Aufrufe) | - | - |
| Kosten für Verschiebung | 0 | 0 |
| Speicherkosten | \$1,254.40 | \$1,684.48 |
| Gesamtkosten | \$2,938.88 | |

| Lebenszyklus-Konfiguration für die Verwendung von verschiedenen Arten von Speichern | | | | |
|--|-----------------------------|------------------------------|-------------------------------|------------|
| | S3 Standard (erste 51200GB) | S3 Standard (Nächste 450 TB) | S3 Standard Infrequent Access | S3 Glacier |
| Speicherplatz in GB | 51200 | 10240 | 24576 | 36864 |
| Preis pro GB | \$0.0245 | \$0.0235 | \$0.0136 | \$0.0045 |
| Speicherverteilung | 42% | 8% | 20% | 30% |
| Anzahl der Dateien | 15059 | 3012 | 7228 | 10842 |
| Übertragungsgebühr (pro 1.000 Aufrufe) | - | - | \$0.0100 | \$0.0360 |
| Kosten für Verschiebung | 0 | 0 | \$0.72 | \$3.90 |
| Speicherkosten | \$1,254.40 | \$240.64 | \$334.23 | \$165.89 |
| Gesamtkosten | | | | \$1,999.79 |

Abbildung 13

Kostenvergleich durch Nutzung von unterschiedlichen Speicherklassen.

Quelle: Eigene Darstellung. S3 Stundensätze: ⁴⁴

Der Punkt wurde als Dezimaltrennzeichen und das Komma als Tausendertrennzeichen verwendet.

5.2.3 Intelligent-Tiering

Intelligent-Tiering verschiebt Dateien auf der Grundlage von Zugriffsmustern. Diese Speicherklasse ist ideal für Daten mit wechselnden oder unbekannten Zugriffsmustern. Wie die Senior Product Manager für S3 Ruhi Dang erklärt, einige Unternehmen haben weder die Zeit noch die finanziellen Möglichkeiten, eine Person einzustellen, die ihre Daten sortiert und in die richtige Speicherklasse einordnet. Intelligent Auto Tiering ist eine attraktive

Lösung für Unternehmen, die jährlich weniger als \$100,000 für Speicher ausgeben ⁴⁵. Abbildung 14 zeigt, wie die Dateien in Abhängigkeit davon, ob auf sie zugegriffen wurde oder nicht, verschoben werden. Wird ein Datei zu einem späteren Zeitpunkt aus der Ebene

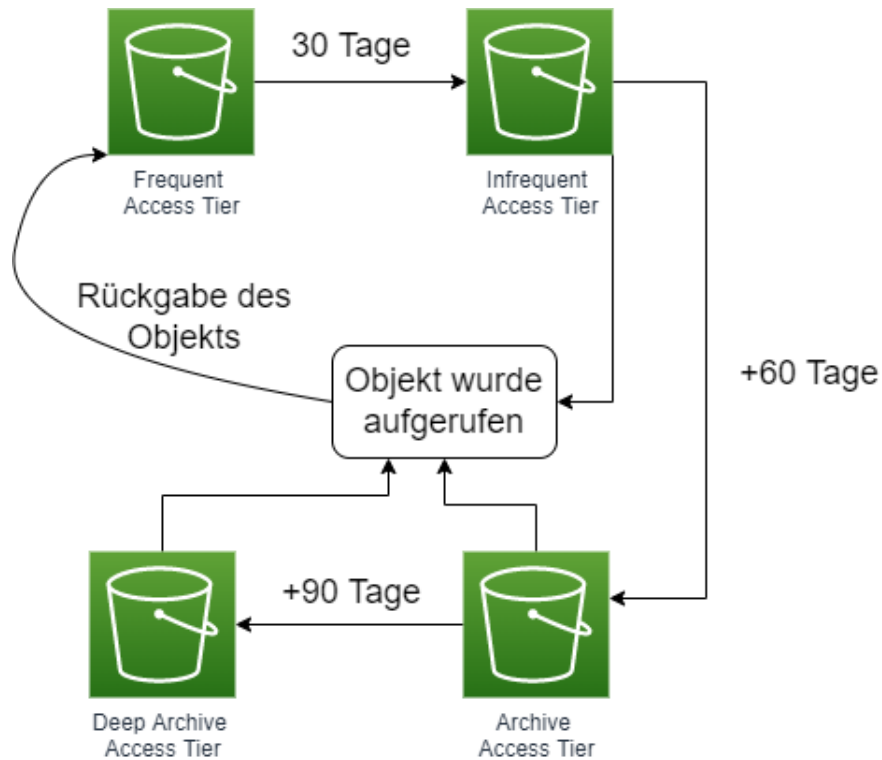


Abbildung 14
Funktionsweise von Intelligent-Tiering
Quelle: Eigene Darstellung

der seltenen Zugriffe aufgerufen, wird es automatisch in eine Speicherklasse der häufigen Zugriffe zurückversetzt.

⁴⁵[17], Minute: 21:12

Zusammenfassung und Ausblick

(To-Do:)

Kapitelweise Kurzdarstellung der Inhalte (inklusive Referenzierung auf die Kapitelnummerierung) => Nach dem Motto: *Was wurde wo beschrieben?*

Kurzdarstellung *Problem – Lösungsweg – Ergebnisse*

Rückkopplung auf die Einleitung: Wurde die Zielstellung der Arbeit und die Fragestellung zufriedenstellend beantwortet?

Kritische Bewertung (sofern nicht bereits im Hauptteil geschehen)

Offene Probleme

Richtung der zukünftigen/möglichen Arbeiten

Erläuterung, warum welche Aspekte in der Arbeit nicht erläutert

Bewusstsein in der gesamten Organisation

Zusätzlich zu den bisher genannten Maßnahmen ist es wichtig, dass Verbraucher von Cloud-Diensten Bewusstsein für die Entstehung von Kosten entwickeln[ODER sensibilisiert werden?]. Von dem Entwickler bis zum der IT-Manager, jeder sollte wissen, dass es so einfach ist, Cloud-Dienste mit ein paar Klicks zu beauftragen⁴⁶. Diese können in kurzer Zeit ungewünschte Kosten verursachen oder sogar über Jahre hinweg wirtschaftliche Schäden verursachen.

Die richtige Personen finden, Ownship verbreiten

Die technischen Maßnahmen zur Überwachung und Kostenreduzierung wurden dargelegt, aber jemand muss diese Analysen, Anpassungen und Entscheidungen durchführen. Deshalb ist es wichtig, bestimmte Personen zu berücksichtigen, die die Verantwortung für das Geschehen in den Cloud-Systemen übernehmen. Idealerweise Menschen, die sich für das Thema interessieren und über die notwendigen Kenntnisse verfügen, um die gesetzten Ziele zu erreichen.

⁴⁶[38], Seite 5

5G/IoT generierte Daten

Mit 5G ist prognostiziert, dass mehr Daten[WIE VIELE / WANN?] automatisch und schnell von Maschinen produziert werden.

Rentabilität bei der Optimierungsmaßnahmen

Kostenoptimierung UND -Überwachung SOLLEN DIE Einsparungen NICHT ÜBERSCHREITEN . TRUSTED ADVISOR NICHT FÜR JEDE FIRMA.

Glossar

Cloud-Computing:

...

Cloud-Dienste:

...

On-Demand:

...

On-Premise:

...

Region:

Die Region ist ein völlig unabhängiges und eigenständiges geografisches Gebiet. Jede Region hat mehrere, physisch getrennte und isolierte Standorte, die als Availability Zones bekannt sind. Beispiele für Regionen sind London, Dublin, Sydney, usw.⁴⁷.

Availability Zone:

Eine Verfügbarkeitszone ist einfach ein Datenzentrum oder eine Sammlung von Datenzentren. Jede Verfügbarkeitszone in einer Region verfügt über eine separate Stromversorgung, Netzwerk und Konnektivität, um die Gefahr eines gleichzeitigen Ausfalls in beiden Zonen zu verringern⁴⁸.

Instance family:

Instanzfamilien sind eine Sammlung von EC2-Instanzen, die nach dem Verhältnis von Speicher, Netzwerkleistung, CPU-Größe und Speicherwerten zueinander gruppiert sind. Zum Beispiel bietet die m4-Familie von EC2 eine ausbalancierte Kombination von Rechen-, Speicher- und Netzwerkressourcen.⁴⁹.

Instagram-Story

⁴⁷[2], Seite 42

⁴⁸[2], Seite 42

⁴⁹[2], Seite 95

Tag

Buckets

PAYG

Metadaten

Startkonfiguration

Scale-In/Out

Quellenverzeichnis

Literatur

- [1] Anders Lisdorf (2021): Cloud Computing Basics: a Non.-Technical Introduction. Apress.
ISBN-13 (pbk): 978-1-4842-6920-6

- [2] AWS Certified Solutions Architect - Associate (SAA-C02)
https://books.google.de/books?id=Dp__DwAAQBAJ&lpg=PA29&ots=T5WqfT25mA&dq=Increase%20efficiencies%3A%20Use%20automation%20to%20reduce%20or%20eliminate%20IT%20management%20activities%20that%20waste%20time%20and%20resources.&pg=PA29#v=onepage&q&f=false
ISBN: 9780137325160
(Abgerufen am 02.11.2021)

Internetquellen

- [1] Accenture Dienstleistungen GmbH. Hohe Erwartungen an die Cloud: Hürden meistern, Mehrwert maximieren
<https://www.accenture.com/de-de/insights/technology/maximize-cloud-value>
(Veröffentlicht am 13.11.2020, abgerufen am 12.04.2021)

- [2] Accenture GmbH: Navigating the barriers to maximizing cloud value (Vollständiger Bericht auf Englisch)
https://www.accenture.com/_acnmedia/PDF-139/Accenture-Cloud-Outcomes-Exec-Summary.pdf#zoom=40
(Veröffentlicht July-August 2020, abgerufen am 29.11.2021)

- [3] AWS Introduction to EC2 Auto Scaling
<https://www.aws.training/Details/Video?id=16387>
(Abgerufen am 23.09.2021)

-
- [4] AWS On-Demand Instances Pricing
<https://aws.amazon.com/de/ec2/pricing/on-demand/>
(Abgerufen am 20.10.2021)
- [5] AWS-Entwicklerzentrum
<https://aws.amazon.com/de/developer/> (Abgerufen am 21.10.2021)
- [6] AWS Entwicklung kostenloser Websites und Webanwendungen
<https://aws.amazon.com/de/free/webapps/> (Abgerufen am 21.10.2021)
- [7] AWS S3 Intelligent-Tiering Adds Archive Access Tiers
<https://aws.amazon.com/de/blogs/aws/s3-intelligent-tiering-adds-archive-access-tiers#:~:text=What%20is%20S3%20Intelligent%2DTiering>
(Veröffentlicht am 09.11.2020)
- [8] AWS Reserved Instances Pricing
<https://aws.amazon.com/de/ec2/pricing/reserved-instances/>
(Abgerufen am 22.10.2021)
- [9] AWS für Amazon EC2 Spot Instances
<https://aws.amazon.com/de/ec2/spot/pricing/> (Abgerufen am 25.10.2021)
- [10] AWS S3 Pricing
<https://aws.amazon.com/de/s3/pricing/> (Abgerufen am 25.10.2021)
- [11] AWS Databases
<https://aws.amazon.com/de/products/databases/learn/>
(Abgerufen am 28.10.2021)
- [12] AWS Saving Plans Pricing
<https://aws.amazon.com/de/savingsplans/compute-pricing/>
(Abgerufen am 02.11.2021)
- [13] AWS Cloud Watch Features
<https://aws.amazon.com/de/cloudwatch/features/> (Abgerufen am 03.11.2021)
- [14] AWS Cloud Watch Events: User Guide
<https://docs.aws.amazon.com/AmazonCloudWatch/latest/events/cwe-ug.pdf#WhatIsCloudWatchEvents> (Abgerufen am 04.11.2021)

-
- [15] AWS Cloud Watch : User Guide
https://docs.aws.amazon.com/AmazonCloudWatch/latest/monitoring/acw-ug.pdf#CloudWatch_Automatic_Dashboards_Focus_Service
(Abgerufen am 04.11.2021)
- [16] AWS Cloud Watch F.A.Q.
<https://aws.amazon.com/de/cloudwatch/faqs/> (Abgerufen am 07.11.2021)
- [17] AWS re:Invent 2019: Guidelines and design patterns for optimizing cost in Amazon S3
<https://youtu.be/UPzsRk2lFWE?t=1279> (Abgerufen am 18.11.2021)
- [18] AWS Pricing Calculator
<https://calculator.aws/#/createCalculator/EC2>
(Abgerufen am 23.11.2021)
- [19] Amazon Simple Storage Service - User Guide
<https://docs.aws.amazon.com/AmazonS3/latest/userguide/s3-userguide.pdf#lifecycle-transition-general-considerations>
(Abgerufen am 24.11.2021)
- [20] Amazon EC2-Spot-Instances
<https://aws.amazon.com/de/ec2/spot/?cards.sort-by=item.additionalFields.startDateTime&cards.sort-order=asc>
(Abgerufen am 26.11.2021)
- [21] AWS Trusted Advisor
<https://aws.amazon.com/de/premiumsupport/technology/trusted-advisor/>
(Abgerufen am 26.11.2021)
- [22] AWS Cost Explorer
<https://aws.amazon.com/de/aws-cost-management/aws-cost-explorer/>
(Abgerufen am 26.11.2021)
- [23] AWS Cost Management Pricing
<https://aws.amazon.com/de/aws-cost-management/pricing/>
(Abgerufen am 30.11.2021)

-
- [24] Amazon EC2 Reserved Instance Marketplace
<https://aws.amazon.com/de/ec2/purchasing-options/reserved-instances/marketplace/>
(Abgerufen am 30.11.2021 - Veröffentlicht: 13.05.2020)
- [25] AWS by Ben Peven: Running Web Applications on Amazon EC2 Spot Instances
<https://aws.amazon.com/de/blogs/compute/running-web-applications-on-amazon-e>
(Abgerufen am 01.12.2021)
- [26] AWS EC2 Spot Instanzen-Anfragen und Preisverlauf
<https://console.aws.amazon.com/ec2sp/v1/spot/home?>
(Abgerufen am 01.12.2021)
- [27] Amazon Elastic Compute Cloud - Benutzerhandbuch für Linux-Instances
https://docs.aws.amazon.com/de_de/AWSEC2/latest/UserGuide/ec2-ug.pdf#spot-best-practices
(Abgerufen am 01.12.2021)
- [28] Microsoft Customer Story-Walgreens Boots Alliance delivers superior customer service with SAP solutions on Azure
<https://customers.microsoft.com/en-us/story/792289-walgreens-boots-alliance-retailers-azure-sap-migration>
(Veröffentlicht am 10. Juni 2020)
- [29] Bertelsmeier, Birgit (o. J.): Tipps zum Schreiben einer Abschlussarbeit. Fachhochschule Köln-Campus Gummersbach, Institut für Informatik.
<http://lwibs01.gm.fh-koeln.de/blogs/bertelsmeier/files/2008/05/abschlussarbeitsbetreuung.pdf> (Veröffentlicht am 29.10.2013).
- [30] SevDesk: Definition von Budgetplanung
<https://sevdesk.de/lexikon/budgetplanung/#budgetplanung-definition>
(Abgerufen am 28.11.2021)
- [31] Indeed:Cost Control Methods: Definitions and Examples
<https://www.indeed.com/career-advice/career-development/cost-control-methods>
(Abgerufen am 29.11.2021)
- [32] Ubuntu, delivered by Canonical:A business guide to hybrid/multi-cloud
https://ubuntu.com/engage/multi-cloud-business-guide?utm_source=

-
- [40] Stern, Adam, The Truth About Cloud Pricing
<https://www.forbes.com/sites/forbestechcouncil/2018/11/16/the-truth-about-cloud-pricing/?sh=1f37bba42f33>
(Veröffentlicht am 16.11.2018)
- [41] Spot by NetApp, What are AWS spot instances?
<https://spot.io/what-are-ec2-spot-instances/>
(Abgerufen am 01.12.2021)
- [42] Putting a Finger on Our Phone Obsession
https://blog.dscout.com/mobile-touches?_ga=2.18241977.1010253397.1637068725-1707869761.1637068725 (Abgerufen am 16.11.2021)
- [43] Statista: 2020 überholt die Cloud lokale Speichermedien
<https://de.statista.com/infografik/18231/cloud-vs-lokaler-speicher/>
(Abgerufen am 18.11.2021)
- [44] Statista: Wie schätzen Sie die Bedeutung Cloud-basierter Anwendungen in Ihrem Unternehmen ein?
<https://de.statista.com/statistik/daten/studie/1221723/umfrage/umfrage-zur-bedeutung-cloud-basierter-anwendungen-im-handel/> (Abgerufen am 25.11.2021)
- [45] Statista: Corona-Krise: Anteile der Unternehmen mit geplanten Veränderungen im Arbeitsalltag nach Arbeitsbereichen in Deutschland im 2. Quartal 2020
<https://de.statista.com/statistik/daten/studie/1140069/umfrage/corona-krise-veraenderungen-im-arbeitsalltag/> (Abgerufen am 25.11.2021)
- [46] Statista: Cloud infrastructure services vendor market share worldwide from 4th quarter 2017 to 3rd quarter 2021
<https://www.statista.com/statistics/967365/worldwide-cloud-infrastructure-services-market-share-vendor/> (Abgerufen am 25.11.2021)
- [47] Statista: Wie viel planen Sie am Black Friday / Cyber Monday auszugeben?
<https://de.statista.com/statistik/daten/studie/1074692/umfrage/hoehe-der-geplanten-ausgaben-am-black-friday-und-cyber-monday-in-deutschland/>
(Abgerufen am 29.11.2021)

A Anhang

(To-Do:)

A.1 Anhang X

Erklärung über die selbständige Abfassung der Arbeit

Ich versichere, die von mir vorgelegte Arbeit selbständig verfasst zu haben. Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten oder nicht veröffentlichten Arbeiten anderer entnommen sind, habe ich als entnommen kenntlich gemacht.

Sämtliche Quellen und Hilfsmittel, die ich für die Arbeit benutzt habe, sind angegeben. Die Arbeit hat mit gleichem Inhalt bzw. in wesentlichen Teilen noch keiner anderen Prüfungsbehörde vorgelegen.

(Ort, Datum, Unterschrift)

Hinweise zur obigen *Erklärung*

- Bitte verwenden Sie nur die Erklärung, die Ihnen Ihr **Prüfungsservice** vorgibt. Ansonsten könnte es passieren, dass Ihre Abschlussarbeit nicht angenommen wird. Fragen Sie im Zweifelsfall bei Ihrem Prüfungsservice nach.
- Sie müssen **alle abzugebende Exemplare** Ihrer Abschlussarbeit unterzeichnen. Sonst wird die Abschlussarbeit nicht akzeptiert.
- Ein **Verstoß** gegen die unterzeichnete *Erklärung* kann u. a. die Aberkennung Ihres akademischen Titels zur Folge haben.