

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, 20 Dezember 2021

Abstract

Zusammenfassung

[Rev]In dieser Arbeit werden Werkzeuge und Maßnahmen untersucht, die zur Kostenkontrolle von AWS-Diensten beitragen. Darüber hinaus werden allgemeine Optimierungsmaßnahmen aufgezeigt, die bereits über (die Jahre hinweg/ mehrere Jahre) von anderen Cloud-Nutzern getestet wurden und von Amazon Web Services (als Best Practices) empfohlen werden. Die Grundlage dieser Untersuchung sind Empfehlungen von Cloud-Anbietern bezüglich Kostenüberwachung und -optimierung, Erfahrungen von Experten dieses Fachgebiets und aktuelle Fachliteratur.

Es ist besonders interessant für Teams, die AWS-Cloud-Dienste in aktuellen Projekten nutzen und ihre Kosten in der Cloud besser verstehen und optimieren möchten. Wenn die Kosten für Cloud-Dienste wie alle anderen Kosten betrachtet werden, ist es konsequent, über ihre Überwachung, Kontrolle und Optimierung nachzudenken. Ein häufiges Problem im Unternehmen ist das fehlende Verständnis der in der Cloud anfallenden Kosten¹. Dieses entzieht die Kontrolle über die Kosten von Cloud-Diensten. Aus diesem Grund stehen Unternehmen, die noch eine 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 cloudbasierten Modellen profitieren möchten.

¹Vgl. Stern, Adam: The Truth About Cloud Pricing, 2018, o.S. [65]

Abstract

Platz für das englische Abstract....

Inhaltsverzeichnis

Abstract	1
Abbildungsverzeichnis	5
Glossar	6
Abkürzungsverzeichnis	9
1 Einleitung	10
1.1 Motivation	10
1.2 Problemstellung	11
1.3 Zielsetzung	12
1.4 Struktur der Arbeit	12
2 Grundlagen	13
2.1 Cloud Economics	13
2.1.1 Skalierbarkeit	14
2.1.2 Flexibilität	14
2.1.3 Selbstbedienung	15
2.1.4 Keine Vorabkosten	15
2.1.5 Technische Fachkompetenz	15
2.2 Amazon Cloud-Dienste	16
3 Zahlungsmodelle	18
3.1 On-Demand-Instanzen	18
3.2 Reservierte Instanzen und Saving Plans	19
3.3 Spot Instanzen	21
3.4 Amazon EC2 Fleet[rev]	22
3.5 Anwendungsfall: TrueCar[rev]	24
4 Kostenüberwachung	28
4.1 AWS CloudWatch	31
4.2 AWS Cost-Explorer	35
4.3 AWS Trusted Advisor[Rev]	38
4.4 Überwachungswerkzeuge gemäß ihrer Verwendung	42

5	Optimierungsmaßnahmen	43
5.1	EC2 Auto Scaling	43
5.1.1	Zeitgesteuerte Skalierung	45
5.1.2	Dynamisches Auto Scaling[Rev]	47
5.1.3	Manual Scaling	48
5.1.4	Predictive Scaling	48
5.2	S3 Optimierung	48
5.2.1	Die richtige Speicherklassen wählen[Rev]	49
5.2.2	Lebenszyklus-Konfiguration	49
5.2.3	Anwendungsbeispiel für eine Lebenszyklus-Konfiguration	51
5.2.4	Intelligent-Tiering	53
	Zusammenfassung und Ausblick	57
	Bewusstsein in der gesamten Organisation	59
	Die richtige Personen finden, Owneship verbreiten	59
	5G is coming	60
	Rentabilität bei der Optimierungsmaßnahmen	60
	Quellenverzeichnis	61
	Anhang	71
I	Vorlage für einer Fakturierungsalarme in CloudWatch	71
II	Alarm für die monatliche Kosten anhand eines Budgets	73
	Erklärung über die selbständige Abfassung der Arbeit	74

Abbildungsverzeichnis

1	Beispiel für ein Tag	7
2	2020 überholt die Cloud lokale Speichermedien	17
3	On-Demand Preise für Amazon EC2	19
4	Mögliche Einsparungen bei reservierten Instanzen and Saving Plans laut AWS	20
5	Mögliche Einsparungen durch Vorauszahlungen	21
6	Monatliche Kosten für eine On-Demand-Instanz im Vergleich zu einer reservierten Instanz	25
7	Vergleich der Zahlungsmodelle	26
8	Trennung der Kosten durch Tags	30
9	Dashboard-Test in CloudWatch	33
10	Kosten nach Projektphasen	34
11	Dashboard mit EC2 und S3 Metriken	37
12	Operationen an Cloud-Diensten in CloudWatch	37
13	AWS Trusted Advisor Kategorien	39
14	Überwachungswerkzeuge gemäß ihrer Verwendung	42
15	Ungenutzte Rechenkapazität ohne automatische Skalierung	43
16	Auto-Scaling-Gruppe nach den Anzahl der Instanzen und Umleitung der Datenverkehr durch dem Application Load Balancer	44
17	Berechnung für ein nicht produktives Umgebung mit Zeitgesteuerte Skalierung	46
18	Nutzung von Tinder, OkCupid und Netflix pro Stunde	47
19	Kostenvergleich durch Nutzung von unterschiedlichen Speicherklassen	52
20	Funktionsweise von Intelligent-Tiering	53
21	Berechnung für die Verwaltung von 120 TB mit AWS Pricing-Calculator für S3 Intelligent-Tiering(1)	54
22	Berechnung für die Verwaltung von 120 TB mit AWS Pricing-Calculator für S3 Intelligent-Tiering(2)	55
23	Budgetalarm	73

Glossar

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 verringern ².

Buckets

Buckets sind in AWS-S3 Behälter, wo Dateien wie Bilder oder Videos gespeichert werden ³.

Cloud-Computing

Das NIST definiert Cloud Computing als das Modell zur Ermöglichung eines allgegenwärtigen, bequemen und bedarfsgerechten Netzzugangs zu einem gemeinsamen Pool konfigurierbarer Rechenressourcen (z. B. Netze, Server, Speicher, Anwendungen und Dienste), die mit minimalem Verwaltungsaufwand oder minimaler Interaktion mit dem Dienstanbieter schnell bereitgestellt und freigegeben werden können ⁴.

Cloud-Dienst

Bei Cloud-Dienste geht es um sämtliche Infrastruktur-Komponenten wie die Server, Rechenleistung, Netzkapazitäten, Kommunikationsgeräte, Speicher, Archivierungs- und Backup-Systeme und andere Komponenten der Rechenzentrum- und Netzinfrastruktur, die von dem Cloud-Service-Provider zur Verfügung gestellt werden. Der Anwender kann über das Netzwerk (i. d. R. Internet) auf die virtuellen Services zugreifen. Beispiele für Cloud-Dienste stellen die Elastic Compute Cloud (EC2) von Amazon, die Microsoft Windows Azure virtuelle Maschinen und die Google Compute Engine ⁵.

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-

²Vgl. AWS Certified Solutions Architect - Associate (SAA-C02), S.42.[1]

³Vgl. AWS: Amazon Simple Storage Service - User Guide. S.4.[20]

⁴Vgl. The NIST Definition of Cloud Computing. S.6 [47]

⁵Vgl. Helmut Krcmar. (2015). Einsatzfelder und Herausforderungen des Informationsmanagements. Informationsmanagement. 6. Auflage S.724[5]

, Speicher- und Netzwerkressourcen⁶.

Instagram-Story

Bei Instagram Stories handelt es sich um kurzen visuellen Content in der Regel Bilder oder kurze Videos, die nach 24 Stunden automatisch aus der Applikation Instagram verschwinden(Stand November 2021)⁷. **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.⁸.

Tag

Ein *Tag* (Markierung) ist eine Markierung, die einer AWS-Ressource zuordnet. Jeder Tag (Markierung) besteht aus einem Schlüssel und einem optionalen Wert⁹.

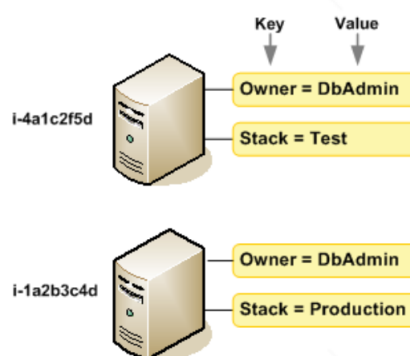


Abbildung 1
Beispiel für ein Tag[28], S.1570.

Metadaten

Metadaten liefern Informationen über den Inhalt eines bestimmten Objekts. Ein Bild kann beispielsweise Metadaten enthalten, die beschreiben, wie groß das Bild ist, die Farbtiefe, die Bildauflösung, wann das Bild erstellt wurde und andere Daten. Die Metadaten eines Textdokuments können Informationen darüber enthalten, wie lang das Dokument ist, wer der Autor ist, wann das Dokument geschrieben wurde und eine kurze Zusammenfassung

⁶Vgl. AWS Certified Solutions Architect - Associate (SAA-C02). S.95[1]

⁷Vgl. Online Marketing: Definition von Instagram Story? o.S.[50]

⁸Vgl. AWS Certified Solutions Architect - Associate (SAA-C02)[1], S.42

⁹Vgl. Amazon Elastic Compute Cloud - Benutzerhandbuch für Linux-Instances[28], S.1570

des Dokuments¹⁰.

Startkonfiguration

Eine Startkonfiguration ist eine Instance-Konfigurationsvorlage, die eine Auto-Scaling-Gruppe zum Starten von EC2-Instances verwendet¹¹.

Maschinelles Lernen

Maschinelles Lernen ist ein Teilbereich der künstlichen Intelligenz (KI). Beim maschinellen Lernen werden Algorithmen darauf trainiert, Muster und Korrelationen in großen Datensätzen zu finden und auf Basis dieser Analyse die besten Entscheidungen und Vorhersagen zu treffen. Auf diese Weise wird die Rechenkapazität von EC2-Instanzen auf der Grundlage früherer Muster vorhergesagt¹².

YAML

YAML ist eine benutzerfreundliche Daten-Serialisierungs Sprache für alle Programmiersprachen¹³.

¹⁰Vgl. Techterms Definition Metadata[60]

¹¹Vgl. Amazon EC2 Auto Scaling - Benutzerhandbuch. S.54 [33]

¹²Vgl. SAP: Definition von maschinellen Lernen[58]

¹³Vgl. YAML Org: Definition von YAML[75]

Abkürzungsverzeichnis

AWS Amazon Web Services

API Application Programming Interface

ASG Auto Scaling Group

CI/CD Continuous Integration / Continuous Deployment

EC Elastic Compute

GK Anzahl der gewonnenen Kunden

GB Gigabyte

JSON JavaScript Object Notation

TCO Total Cost of Ownership

MK Anfallende Marketingkosten

PAYG Pay-as-you-go

SSO Single Sign-On

TB Terabyte

VK Vertriebskosten

KPI Key Performance Indicators

1 Einleitung

1.1 Motivation

[Mein persönliches Statement: Die Möglichkeit, eine IT-Infrastruktur von zu Hause aus einzurichten, hat mein Interesse am Cloud-Computing geweckt, aber die Notwendigkeit, meine Kreditkarte einzugeben, ohne zu wissen, wie viel davon abgebucht wird, hat mich ziemlich verunsichert. (Wie kann der Satz davor mit dem kommenden? ODER lieber wo anderes der erste Satz plazieren?/MARCEL:Am Ende der Motivation scheint es besser zu klingen)] Die zunehmende Digitalisierung von Geschäftsmodellen, die auch durch die Corona-Pandemie vorangetrieben wird, lässt Cloud-basierte Applikationen an Bedeutung gewinnen.¹⁴ Als direkte Folge davon ist die Nachfrage nach Server- und Speicherkapazität gestiegen. Die Relevanz von *Amazon Web Services*, kurz AWS, in dem Bereich der Cloud-Computing ergibt sich aus einer vor kurzem veröffentlichte Studie von Raj Bala et al.. Diese wies eindrücklich daraufhin, dass AWS der aktuell weltweit führende Cloud-Anbieter anhand ihrer Klassifikation (*Magic Quadrant*¹⁵) für Cloud-Infrastruktur und Plattform-Services sei (Bala et al, 2021, o.S.,[52]). AWS erscheint nicht nur aus diesem Grund als Fallbeispiel für diese Arbeit passend, weitere bedeutsame Faktoren sind seine frühe Präsenz (2006) als Cloudanbieter und seines großen Angebotes an Cloud-Diensten, welche für zahlreiche Anwendungsfälle geeignet sind.¹⁶

¹⁴Es sei an dieser Stelle darauf hingewiesen, dass in diesem Kontext Ahrens die Bedeutung Cloud-basierter Anwendungen im Bereich von deutschen Handelsunternehmen untersuchte (Vgl. Ahrens 2021)[70], sowie das *ifo Institut* anschaulich die strukturellen Veränderungen von der Corona-Pandemie auf den Arbeitsalltag in Deutschland nachzeichnete (Vgl. ifo Institut 2020)[69].

¹⁵Vgl. Laut Gartner stellt der Magic Quadrant eine zweidimensionale Matrix mit vier Quadranten dar. Jeder Quadrant steht für einen Unternehmenstypus im Markt. Im Uhrzeigersinn von links unten beginnend sind dies: *Nischenanbieter*, *Herausforderer*, *Marktführer* und *Visionäre*

¹⁶Die aktuellen Marktführer im Bereich der *Cloud-Computing* weltweit sind AWS, Google, Telekom und Microsoft (Vgl. Synergy Research Group 2019, o.S.[73])

1.2 Problemstellung

Adam Stern wies in dem *Forbes*-Magazin daraufhin, dass ungefähr die Hälfte der US-amerikanischen Unternehmen Schwierigkeiten hätten ihre Kosten zu begründen (Stern 2018, o.S.).

„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.“¹⁷

Darüber hinaus weist Tobias Regenguß und Jochen Malinowski von Accenture GmbH in einer Untersuchung, dass es den Unternehmen an fachlichem Know-How in Cloud-Computing mangelte. Diese stelle eine der größten Hindernisse dar, um einen Wechsel von On-Premise- zu Cloud-basierten Systemen gewährleisten zu können¹⁸.

Kostenoptimierung für Cloud-Dienste ist ein wichtiger Punkt, da man ohne Optimierungsmaßnahmen mit höheren Kosten rechnen müsse als bei On-Premise Systemen (Anders Lisdorf¹⁹).

”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.”²⁰

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. Außerdem sollte untersucht werden, wie mit der richtigen Auswahl an Diensten Kosten optimiert werden. Es wird untersucht, welche Maßnahmen nötig sind, um unerwartet hohe Kosten bei Cloud-Diensten zu vermeiden. Darüber hinaus wird untersucht wie die Kosten von Cloud-Diensten minimiert beziehungsweise optimiert werden können. Diese Arbeit untersucht speziell die Kostenoptimierung von Amazon S3-Speichereinheiten und EC2-Server-Instanzen mithilfe von folgenden Überwachungswerkzeuge: Cost-Explorer, CloudWatch und Trusted Advisor.

¹⁷Stern, Adam, 2018, The Truth About Cloud Pricing.[65]

¹⁸Vgl. Regenguß und Malinowski, Hohe Erwartungen an die Cloud: Hürden meistern, Mehrwert maximieren. 2020 o.S.(Webversion) oder S.11 in der PDF-Version auf Englisch[1]

¹⁹Vgl. Anders Lisdorf. Cloud Computing Basics: a Non-Technical Introduction. S.152. [4]

²⁰Anders, Lisdorf, 2021, S.152.[4]

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 Amazon 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 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 (TCO²¹)²².

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 Fachkraft²³. 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 die Nutzerin oder der Nutzer von Cloud-Diensten Verantwortung für die anfallenden Kosten übernimmt.

²¹Vgl. TCO: Total Cost if Ownership

²²Vgl. Ubuntu, delivered by Canonical: A business guide to hybrid/multi-cloud, S.2.[46]

²³Vgl. Larry Carvalho and Matthew Marden from IDC: Quantifying the Business Value of Amazon Web Services, 2015, S.1[48]

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 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.²⁴

Dies war der Fall bei Walgreens 2020 in den Vereinigte 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.,”²⁵

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].

²⁴Vgl. AWS Certified Solutions Architect - Associate (SAA-C02), S.29.[1]

²⁵Vgl. Microsoft Customer Story-Walgreens Boots Alliance delivers superior customer service with SAP solutions on Azure.[41]

- Vielfältige Hardware-Konfigurationen.

Durch die Vielzahl der verfügbaren Diensten ist es möglich, Prototypen und Experimente in kurzer Zeit durchzuführen²⁶. 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[Rev] 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 werden²⁷. 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 angeboten²⁸. Dies erfordert keine Vorauszahlungen für die Nutzung von vielen 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

²⁶Vgl. IDC Business Value of AWS 2015 S.7[48]

²⁷Vgl. Cloud Computing Basics: a Non.-Technical Introduction, S.28[4]

²⁸Vgl. Die aktuellen Marktführer im Bereich der *Cloud-Computing* weltweit sind AWS, Google, Telekom und Microsoft (Vgl. Synergy Research Group 2019, o.S.[73])

bei der Migration in die Cloud konfrontiert werden. In der von Accenture im Jahr 2020 durchgeführten Umfrage gaben 38% der Befragten an, dass fehlende Kompetenzen im Unternehmen im Bezug auf die Cloud ein Hindernis für eine Cloud-Migration ist²⁹.

2.2 Amazon Cloud-Dienste

Von dieser Stelle der Arbeit an liegt der Fokus auf den Cloud-Diensten von Amazon Web Services, die als *AWS-Dienste* bezeichnet werden. Einer der am häufigsten genutzten AWS-Dienste ist *Amazon Elastic Computing Instances EC2*, mit dem virtuelle Maschinen erstellt werden können³⁰. Ein weiterer wichtiger AWS-Dienst ist *Amazon Simple Storage Service (S3)*, der zum Speichern von Objekten³¹ verwendet wird. Amazon Elastic Computing Instances EC2 werden im Folgenden in dieser Arbeit als *EC2-Instanzen* und Amazon Simple Storage Service als *Amazon S3* oder *Amazon S3-Speichereinheiten* bezeichnet.

Wie Lynn Langit, eine erfahrene Cloud-Architektin, feststellt, können bis zu 80% der AWS-Rechnung aus Gebühren für EC2-Instanzen bestehen³². Laut des AWS Solutions Architekten Daniel Peña Silva³³ ist Amazon S3 einer der am häufigsten genutzten AWS-Dienste. Deshalb konzentrieren sich in dieser Arbeit die Überwachungs- und Optimierungsmaßnahmen hauptsächlich auf EC2-Instanzen und Amazon S3.

Wie in Abbildung 2 zu sehen ist, werden darüber hinaus seit 2020 weltweit mehr Daten in Serverfarmen als auf lokalen Geräten gespeichert³⁴. 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 Untersuchung sprengen würde.

²⁹Vgl. Accenture Dienstleistungen GmbH. Hohe Erwartungen an die Cloud: Hürden meistern, Mehrwert maximieren, S.11[1]

³⁰Vgl. Cloud infrastructure services vendor market share worldwide from 4th quarter 2017 to 3rd quarter 2021.[71]

³¹Vgl. Objekte sind in AWS die Grundeinheit in welchen Dateien in den Amazon S3-Speichereinheiten gespeichert werden. Neben den Objekten werden Metadaten, wie das Datum der Objekterstellung und das Datum der letzten Aktualisierung gespeichert.

³²Vgl. Lynn Langit, LinkedIn Learning: AWS Controlling Cost. Minute 0:20-0:45[57]

³³Vgl. Daniel Peña Silva, LinkedIn: Listado de todos los Servicios de AWS.[56]

³⁴Vgl. Statista: 2020 überholt die Cloud lokale Speichermedien.[68]



Abbildung 2
Statista: 2020 überholt die Cloud lokale Speichermedien [68]

Dieses grundlegende Kapitel hat einige potenzielle Vorteile der Nutzung von Cloud-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 *reservierte Instanzen* (*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 AWS-Diensten, 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 AWS-Website in der Sektion EC2 - On-Demand-Preise³⁵ zu finden. In der Abbildung 3 werden Preisbeispiele für die Region Ohio verfügbaren Linux-Instanzen gezeigt.

³⁵Vgl. AWS On-Demand Instances Pricing.[4]

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 3
On-Demand Preisbeispiele von EC2-Instanzen ³⁶

Es ist zu beachten, dass Instanzen mit denselben Eigenschaften (Instanzfamilie, Arbeitsspeicher, Netzwerkleistung usw.), 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 USD/Stunden³⁷ gemessen wird. Um die reduzierten Preise zu bekommen, müssen Verträge über ein oder drei Jahre abgeschlossen werden.

Die Abbildung 4 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.

³⁷Vgl. AWS-Services werden in US-Dollar abgerechnet. Zahlungen in anderen Währungen sind auch möglich. Quelle: AWS-Console in Kontoeinstellungen.

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 4

Mögliche Einsparungen bei reservierten Instanzen and Saving Plans laut AWS ³⁸

Compute Saving Plans³⁹ bieten die Flexibilität EC2-Instanzen nach Familie⁴⁰, Größe, Verfügbarkeitszone (AZ), Betriebssystem oder Mandant zu wechseln. Diese Option ist bei 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.“ ⁴¹

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* (Marktplatz für den Kauf von reservierten Instanzen) zur Verfügung gestellt werden. Somit kann ein Teil der Investition zurückgeholt werden. Dies ist für Standard reservierten Instanzen möglich. Diese Instanzen werden in Spot-Instanzen umgewandelt, damit andere Nutzer sie beantragen können. Dafür sollte der Instanz-Anbieter eine Servicegebühr in Betracht ziehen. Stand November 2021 beträgt diese Gebühr 12%⁴².

³⁹Vgl. AWS Saving Plans Pricing[12].

⁴⁰Vgl. AWS Certified Solutions Architect - Associate (SAA-C02), S.95.[1].

⁴¹Vgl. AWS Saving Plans Pricing[12].

⁴²Vgl. Amazon EC2 Reserved Instance Marketplace[25].

Optionale Vorauszahlung

Zusätzlich gibt es bei Saving Plans und reservierten Instanzen die Option im Voraus zu zahlen. Im Gegenzug wird ein niedrigerer Gesamtpreis angeboten. AWS bietet drei verschiedene Optionen an. Diese sind eine teilweise, keine oder eine vollständige Vorauszahlung⁴³. Bei teilweiser Vorauszahlung ist eine Anzahlung von etwa 50% zu leisten.

Die Abbildung 5 zeigt den Vergleich zwischen den drei Optionen für Vorauszahlungen. Hier wird deutlich, dass es kaum einen Unterschied zwischen einer teilweisen Vorauszahlung und keiner 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 5

Mögliche Einsparungen durch Vorauszahlungen für EC2 Instanzen in Saving Plans
Zahlungsmodell

Eigene Darstellung. Quelle: AWS Pricing Calculator[18].

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.

⁴³Vgl. AWS Pricing Calculator[18].

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 ?? 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-Flotten oder auf Englisch *fleet of instances*, bieten bei AWS die Möglichkeit mehrere Spot-Instanzen anzufordern, um einen bestimmten Bedarf an Rechenleistung zu decken⁴⁴. Spot-Instanzen können auch für produktive Umgebungen verwendet werden⁴⁵. 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, dass Spot-Instanzen unterbrochen werden können.

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

Wahl der Spot-Instanzen

Die zu berücksichtigenden Instanzen für die Instanzen-Flotte, müssen den Anforderungen der Applikation entsprechen. Um die Wahrscheinlichkeit zu erhöhen, dass mehr Spot-Instanzen gefunden werden, ist es empfehlenswert, die Kriterien der Suche zu erweitern. Dies kann erreicht werden, indem Instanzen ähnlicher Typen einbezogen werden. Die Berücksichtigung von Instanzen von Familien mit mehr Leistung als erforderlich, ist ebenfalls eine gute Option[26], da der Preis für Spot-Instanzen trotz höherer Leistung geringer sein wird als bei einem On-Demand Zahlungsmodell.

⁴⁴Vgl. Amazon Elastic Compute Cloud - Benutzerhandbuch für Linux-Instances, S.708[28].

⁴⁵Vgl. Running Web Applications on Amazon EC2 Spot Instances[26].

Maximaler Stundenpreis

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, da sie keinen starken Preisschwankungen unterliegen. Die aktuellen Preis und der Preisverlauf von Spot-Instanzen können in auf der AWS-Konsole⁴⁶ abgefragt werden. Diese Informationen sind nur mit einem AWS-Konto zugänglich.

Festlegung von On-Demand-Anteil

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-Instanzen⁴⁷. Im Fall von vorhandenen reservierten Instanzen oder Instanzen von Saving Plans werden On-Demand-Instanzen zum entsprechend reduzierten Preis berechnet⁴⁸.

Auto Scaling Groups

Auch als *EC2-Auto-Scaling-Gruppe*(ASG) bezeichnet, ist diese für die Skalierung der zu startenden Instanzen verantwortlich. Dazu wird eine Startkonfiguration benötigt, welche definiert, unter welchen Bedingungen Instanzen gestartet oder beendet werden sollen. 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 bezahlen⁴⁹.

⁴⁶Vgl. AWS EC2 Spot Instanzen-Anfragen und Preisverlauf[27].

⁴⁷Vgl. Running Web Applications on Amazon EC2 Spot Instances. o.S.[26]

⁴⁸Vgl. Amazon Elastic Compute Cloud - Benutzerhandbuch für Linux-Instances, S.690[28].

⁴⁹Vgl. Amazon Elastic Compute Cloud - Benutzerhandbuch für Linux-Instances, S.709[28].

3.5 Anwendungsfall: TrueCar[rev]

Instanzen in Zahlungsmodellen, die zu zeitlichen Verpflichtungen führen, bergen die Gefahr, dass die benötigte Rechenkapazität mittel- bis langfristig falsch eingeschätzt wird. Einerseits kann die reservierte Rechnerkapazität zu gering eingeschätzt werden. Als Konsequenz 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. TrueCar Inc. ist eine Preis- und Informations-Website für Neu- und Gebrauchtwagenkäufer mit Sitz in Santa Monica, Kalifornien⁵⁰. Dank einer Optimierungsstrategie konnten sie ihre AWS-Kosten durch die Nutzung reservierter Instanzen um etwa 40% senken⁵¹.

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-Dienste⁵². Nachdem das Team von TrueCar die notwendigen Informationen, insbesondere über die reservierten Instanzen, verstanden haben, wurde die benötigte Rechenkapazität ermittelt. In dem Artikel über TrueCar wurde nicht explizit erläutert, wie die benötigte Rechnerkapazität berechnet wurde. Diese Informationen werden jedoch von Cost-Explorer bereitgestellt. Cost-Explorer bietet die Möglichkeit, die Nutzung der AWS-Diensten 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-Point⁵³ dazwischen zu finden. Der Break-Even-

⁵⁰Vgl. Die Quelle dieser Informationen ist ein Artikel[59], der auf www.medium.com veröffentlicht wurde. Dass der Artikel von TrueCar stammt, wird durch die Tatsache bestätigt, dass deren Website www.truecar.com/who-we-are/ zu dem hier erwähnten Artikel führt.

⁵¹Vgl. How TrueCar Saves 40% on AWS with EC2 Reserved Instances. O.s. [59].

⁵²Vgl. Accenture Dienstleistungen GmbH. Hohe Erwartungen an die Cloud: Hürden meistern, Mehrwert maximieren, S.11[1].

⁵³Vgl. Marceil Schweitzer und Ernst Troßmann: Die Break-Even-Analyse dient als Entscheidungshilfe für das Management. Bei einer Break-even-Analyse geht es immer um eine Gegenüberstellung positiver

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 6 dargestellt wird liegt der Break-Even-Point zwischen dem Monat acht und neun. Im Fall, dass der Verbrauch der Instanzen vor dem Monat auch enden würde, wäre es nicht empfehlenswert Instanzen zu reservieren, sondern nur mit On-Demand Instanzen zu arbeiten. Die Berechnung wurde für den Zeitraum von 1 Jahr durchgeführt. In dem Prozess haben die Entwickler von TrueCar Mitarbeiter von

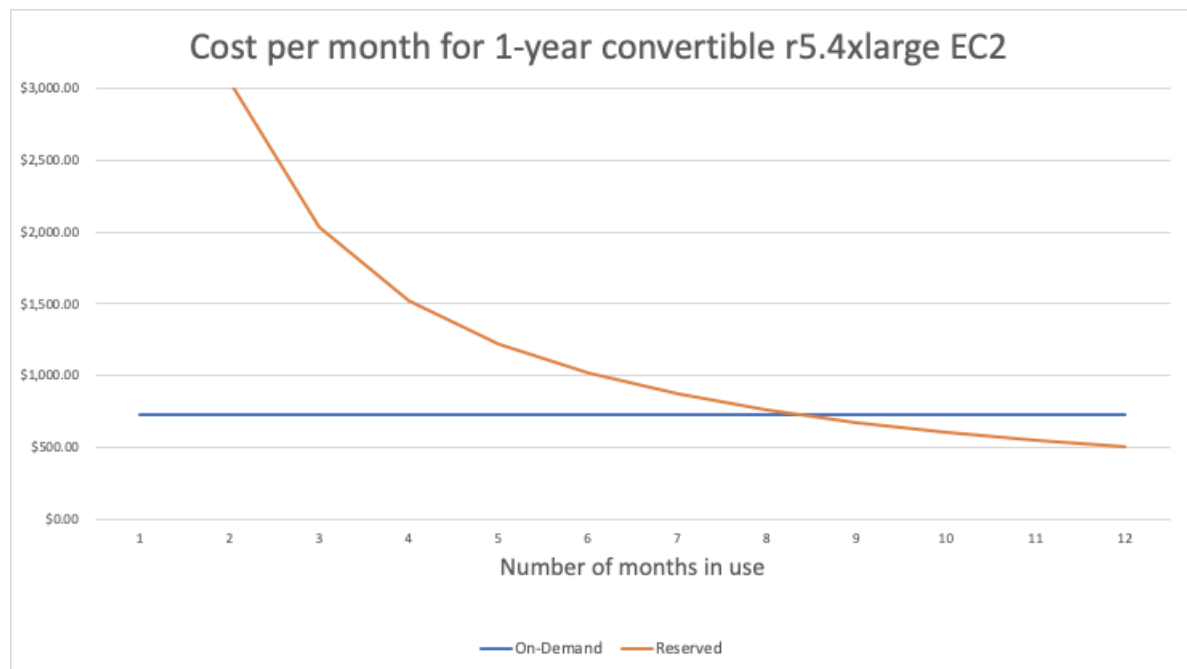


Abbildung 6
Monatliche Kosten für eine On-Demand-Instanz
im Vergleich zu einer reservierten Instanz.

Quelle: Medium: How TrueCar Saves 40% on AWS with EC2 Reserved Instances[59]

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

Mit Cost-Explorer wurden die folgenden zwei Metriken überwacht:

und negativer Wirkungen von Maßnahmen, S.14.[2]

RI-Coverage, welche zeigt, wie viel der On-Demand-Instanzen durch reservierten Instanzen abgedeckt wird. Ziel ist hierbei das *RI-Coverage* (Abdeckung 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. Stattdessen sollte die Nutzung von On-Demand-Instanzen gering gehalten werden.

Um diese Metriken im Blick zu behalten und nicht jeden Tag den Cost-Explorer aufrufen zu müssen, wurde eine Benachrichtigung an *Slack*⁵⁴ eingerichtet. Dies war über die Cost-Explorer API und eine Lambda-Funktion möglich.

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 7
Vergleich der Zahlungsmodelle nach Eigenschaft und Anwendungsfall
Eigene Darstellung. Quelle: [4, 8, 12, 21, 66]
Plusservers: Kostenoptimierung in AWS S.9.[61]

⁵⁴Vgl. Slack ist eine Messaging-App für Unternehmen[64].

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.

Darüber hinaus wurde ein Anwendungsfall vorgestellt, der erhebliche Einsparungen bei der Verwendung reservierter Instanzen zeigt.

Im nächsten Kapitel werden CloudWatch, Cost-Explorer und Trusted Advisor vorgestellt. Diese Werkzeuge sollen ein besseres Verständnis über die Nutzung und Kosten von AWS-Diensten, die Analyse von Metriken ermöglichen und Empfehlungen zur Kostenoptimierung geben.

Für das On-Demand-Zahlungsmodell gibt es keine Kostenreduzierung, aber wie in Unterkapitel 5.1.1 gezeigt gibt Maßnahmen, um die Nutzung von Instanzen zu reduzieren.

4 Kostenüberwachung

Die von Amazon Web Services(AWS) zu Verfügung gestellte Überwachungswerkzeuge werden in diesem Kapitel vorgestellt.

CloudWatch sammelt Metriken von AWS-Diensten und bietet die Möglichkeit, Alarmer und Aktionen zu konfigurieren, die wiederum AWS-Diensten auf der Grundlage dieser Metriken betreffen. Für die Visualisierung von Metriken bietet CloudWatch die Erstellung von personalisierten Dashboards. Cost-Explorer konzentriert sich auf die Überwachung der Nutzung von AWS-Diensten und der dadurch verursachten Kosten. Diese bietet die Möglichkeit Kosten- und Nutzungsberichte der AWS-Diensten zu erstellen. Solche Informationen dienen zugrunde für Budgetierung, Verfolgung von KPIs und Entscheidungsfindung in Bezug auf die operative Planung im Unternehmen. Die vorgenannten Konzepte werden in Unterkapitel 4.2 näher erläutert. Trusted Advisor bietet konkrete Empfehlungen auf der Grundlage von AWS Best-Practices und individuelle Prüfungen von AWS-Diensten.

Es existieren weitere Überwachungswerkzeuge bei AWS, auf die in dieser Arbeit nicht eingegangen wird, weil sie einen anderen Fokus als Kostenüberwachung und -optimierung haben. Zum Beispiel CloudTrail, welches für die Überwachung von Governance, Compliance, Betrieb und Risiken im AWS-Konto ist. Mit CloudTrail können Benutzeraktivitäten über AWS-Dienste durch Ereignisse verfolgt werden⁵⁵.

Ein weiteres Werkzeug ist AWS X-Ray, welches zur Überwachung von Anwendungsleistung verwendet wird. Dies unterstützt Entwickler bei der Analyse und Fehlersuche in verteilten Produktionsanwendungen. Mit X-Ray kann man herausfinden, wie gut Anwendungen und ihnen zugrunde liegenden Dienste funktionieren. Auf diese Weise können Ursache von Leistungsproblemen und Fehlern ermittelt und behoben werden⁵⁶.

⁵⁵Vgl. AWS: CloudTrail User Guide Version 1.0: What Is AWS CloudTrail?, S.1 [29]

⁵⁶Vgl. AWS: X-Ray Developer Guide: What is AWS X-Ray?, S.1[29],

Tag Policies/Tagging-Strategie[Rev]

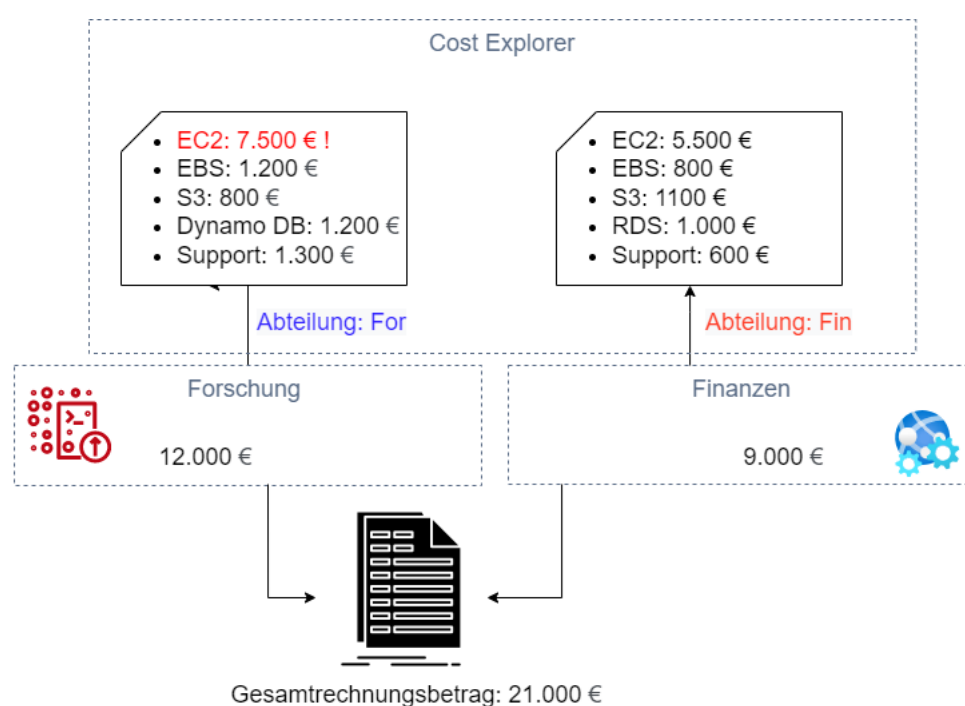
Tags sind bei AWS Information in Form von Metadaten, die an AWS-Dienste zugewiesen werden kann⁵⁷. Ein Tag besteht aus einem Tag-Schlüssel und einem Tag-Wert. Beispiele für Tag-Schlüssel sind Abteilung, Projekt, Team, Region, Art des Dienstes und Umgebung. Tag-Werte für den Tag-Schlüssel *Abteilung* könnten Buchhaltung, Finanz, Entwicklung oder Marketing sein. Sowohl bei Tag-Schlüssel als auch bei Tag-Werte wird zwischen Groß- und Kleinschreibung unterschieden.

Anwendungsbeispiel für Tags und Cost-Explorer

Durch die Verwendung von Tags ist es möglich, die Kosten auf den von der Organisation festgelegten Tags zu verfolgen. 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 AWS generierten Rechnung bemerkbar, um den Grund für diesen Anstieg genauer zu verstehen, muss ihre Ursache untersucht werden. Werkzeuge wie Cost-Explorer zusammen mit einer Tag-Strategie machen diese Art von Analyse möglich.

In der Abbildung 8 wird ein Szenario vorgestellt, wo die Kosten für EC2-Instanzen der Forschungsabteilung kontinuierlich angestiegen sind. Mitarbeiter der Forschungsabteilung waren nicht in der Lage, die Kostensteigerungen zu begründen. Um die von den einzelnen Abteilungen verursachten Kosten zu trennen, wurde ein Tag-Schlüssel mit dem Namen *Abteilung* angelegt. Um anschließend jeder AWS-Dienst einen Tag-Wert entsprechend seiner Abteilung zuzuweisen. Mit Hilfe des Cost-Explorer konnte festgestellt werden, dass die Kosten für EC2 der Forschungsabteilung im Laufe der Zeit gestiegen sind. Nach Angaben der Abteilungsleiter hatte die Nutzung der Rechnerkapazität nicht zugenommen.

⁵⁷Vgl. AWS: AWS – Allgemeine Referenz - Referenzhandbuch. S.681[31]



Monatliche Kosten pro Abteilung			
Monat	Forschung	Finanzen	Gesamtkosten
Mai 2021	€5,600.00	€8,900.00	€14,500.00
Juni 2021	€6,000.00	€8,300.00	€14,300.00
Juli 2021	€7,500.00	€8,000.00	€15,500.00
August 2021	€9,000.00	€9,200.00	€18,200.00
September 2021	€12,000.00	€9,000.00	€21,000.00

Abbildung 8

Trennung der Abteilungskosten durch Tags.

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

Im vorliegenden Fall wurde festgestellt, dass Gastpraktikanten in der Forschungsabteilung Experimente durchführt, in den EC2-Instanzen genutzt wurden. Die Instanzen wurden nach Beendigung des Aufenthalts nicht mehr abgeschaltet und haben kontinuierlich Kosten verursacht. Für diesen hypothetischen Fall wurde die Ursache für den Anstieg der Gesamtkosten einer einfachen Organisation mit zwei Abteilungen und wenigen Cloud-Diensten ermittelt. Es gibt Unternehmen mit viel komplexeren Strukturen als diese, die weitaus mehr Cloud-Dienste in Anspruch nehmen. Für Unternehmen ist eine Tagging-Strategie von Relevanz, um Kosten(Ausgaben⁵⁸) für die Buchhaltungsabteilung genauere

⁵⁸Vgl. Eine Ausgabe im Rechnungswesen liegt beim Abfluss von Zahlungsmitteln und/oder beim Eingehen von Zahlungsverpflichtungen in Form von Geldverbindlichkeiten, z.B. bei der Zahlung von

Daten zu ermittelt und um Budgets auf der Grundlage früherer Projekte erstellen zu können. Die Kostenüberwachung ist mit einer Tag-Strategie auf eine detaillierte Ebene möglich. Je nach festgelegten Tags können detaillierte Analysen der Cloud-Nutzung und -Kosten über Produkte, Einheiten, Umgebungen oder beliebige andere Bereiche hinweg erstellt werden⁵⁹.

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, welche zur Verlaufsanalyse und der Entscheidungsfindung in Bezug auf Optimierung und Fehlerbehebung hilfreich sind. 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 zur Überwachung von EC2-Instanzen in CloudWatch ist die CPU-Auslastung oder *CPU-Utilization* auf Englisch. 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 Kapazitätsbedarfs⁶⁰. 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.

Ereignisse

Ein Ereignis ist in CloudWatch eine Änderung in einem AWS Dienst. AWS-Dienste können Ereignisse erzeugen, wenn sich ihr Status ändert. Beispielsweise, wird ein Ereignis erzeugt,

Dienstleistungen, vor [42].

⁵⁹Vgl. Cloud Computing Basics: a Non.-Technical Introduction. S.152.[4]

⁶⁰Vgl. AWS Certified Solutions Architect - Associate (SAA-C02), S.185.[1]

wenn *Amazon EC2 Auto Scaling*, Instanzen gestartet oder beendet werden⁶¹ oder wenn eine bestimmte Menge an Speicherplatz in einem *Bucket* erreicht wurde. Ein *Bucket* ist ein Behälter, in dem Objekte bei Amazon S3 gespeichert werden⁶². 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 werden⁶³.

Ziele

Ziele oder Targets sind AWS-Dienste, die aufgerufen werden, wenn eine Regel ausgelöst wird. *EC2 instances*, *AWS Lambda functions* und *Amazon SNS*⁶⁴ sind unter anderem mögliche Ziele. Die Ziele einer Regel müssen sich in derselben Region wie die Regel befinden⁶⁵.

Benachrichtigungen

Benachrichtigt zu werden ist wichtig, um relevante Ereignisse nicht zu verpassen und rechtzeitig Maßnahmen zu ergreifen. Mit CloudWatch können Alarmer eingerichtet werden, die durch Metriken wie die CPU-Auslastung und Gebühren von einem spezifischen AWS-Dienst ausgelöst werden. Benachrichtigungen können durch Amazon SNS oder zu einer E-Mail-Adresse geschickt werden.

Zu Testzwecken wurde ein Alarm erstellt, indem eine monatliche Ausgabengrenze von 9 Euro für das AWS-Testkonto für diese Arbeit festgelegt wurde. Dieser ist in Anhang II zu finden.

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 nötig⁶⁶. Die in den Dashboards enthaltenen Informationen sind nicht nur für ihre Autoren von Relevanz. Weitere Personen

⁶¹Vgl. AWS Cloud Watch Events: User Guide. S.1[14]

⁶²Vgl. Amazon Simple Storage Service User Guide, S.4[20]

⁶³Vgl. AWS Cloud Watch : User Guide. S.2[14]

⁶⁴Vgl. Amazon SNS ist ein AWS-Dienst für die Benachrichtigung an Personen und an Applikationen.[32]

⁶⁵Vgl. AWS Cloud Watch Events: User Guide. S.2[14]

⁶⁶Vgl. AWS Cloud Watch : User Guide. S.28[15]

innerhalb oder außerhalb einer Organisation können Zugriff auf Dashboards mit nützlichen Informationen bekommen, um Prozesse zu beschleunigen und Probleme schneller zu beheben. Um den Zugriff auf das Dashboard zu beschränken, ist es möglich, den Zugriff auf bestimmte Personen per E-Mail oder über SSO-Anmeldeinformationen⁶⁷ zu beschränken. Single Sign-On(SSO⁶⁸) ist ein Prozess der einmaligen Authentifizierung und Zugriff auf mehrere Ressourcen. Außerdem hat die Einbindung von Dashboard-Informationen auf Intranet-Portale das Potenzial, Transparenz und eine schnelle Verbreitung von Informationen zu schaffen⁶⁹.

Zu Testzwecken wurde ein Dashboard mit einigen Widgets⁷⁰ erstellt. Das erste Widget in der Abbildung 13 zeigt, wie oft auf die Objekte eines Buckets in S3 zugegriffen wird. Die Anzahl der Aufrufe an der CloudWatch-API wird in dem zweiten Widget gezeigt. Das letzte Widget zeigt die CPU-Auslastung und den eingehenden Netzwerkverkehr von einem Spot-Instanz. Diese Widgets verwenden Standardmetriken, deshalb verursachen sie keine Kosten.

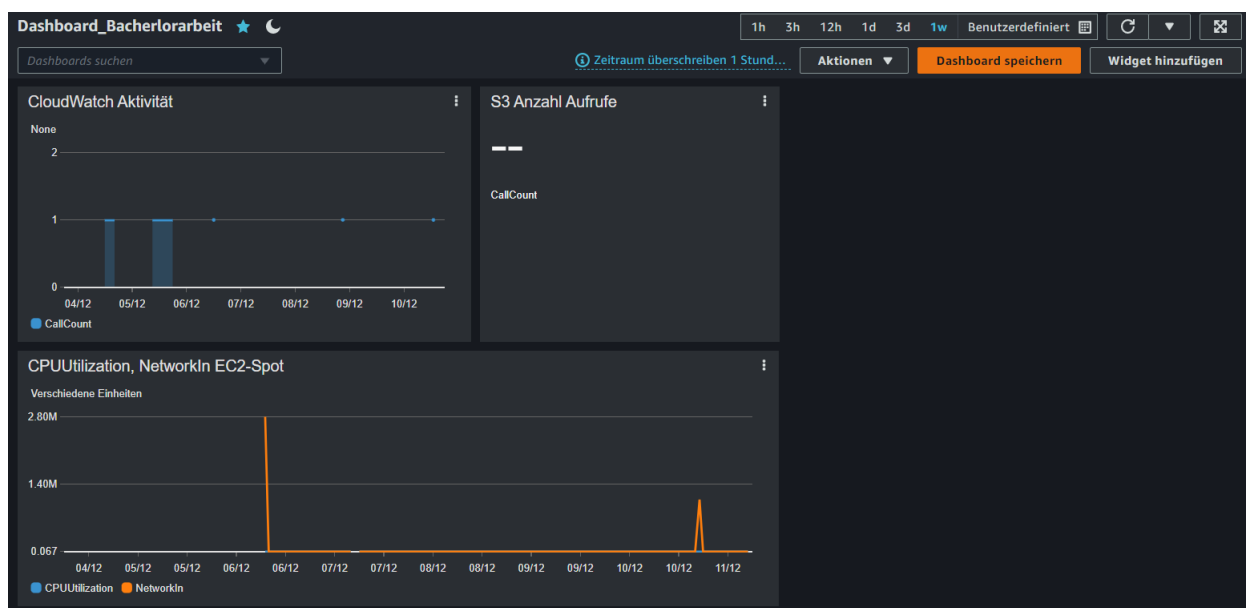


Abbildung 9
Dashboard-Test in CloudWatch.

⁶⁷Vgl. AWS Single Sign-On.[35]

⁶⁸Vgl. Ziel von SSO ist es, die Anzahl von Login und Passwort in heterogenen Umgebungen zu reduzieren. Securing User Authentication using Single SignOn in Cloud Computing[74].

⁶⁹Vgl. Business Knowledge Management: Wertschöpfung durch Wissensportale[3].

⁷⁰Vgl. Ein Widget ist ein grafischer Weg, um Metriken in CloudWatch darzustellen. Unter anderem gibt es Widgets für Zahlen, Linien- und Balkendiagramme.

Wie bereits erwähnt, es ist möglich den Zugriff von Dashboards freizugeben, ohne Zugang zu Ihrem eigenen AWS-Konto gewähren zu müssen. Das hier erwähnte Dashboard wurde für den öffentlichen Zugriff temporär freigegeben. Über den Folgenden Link kann man auf das Dashboard zugreifen: t.ly/fNbyT

Fakturierungsalarme mit CloudWatch

AWS CloudWatch empfängt Abrechnungsmetriken von allen AWS-Diensten. Auch von AWS-Rechnungen, auf der Grundlage dieser Metriken ist es daher möglich, Regeln zu erstellen, die bei Überschreitung des geplanten Budgets Alarmen in Form von Benachrichtigungen auslösen. Wenn ein bestimmter Prozentsatz oder Betrag des festgelegten Budgets überschritten wurde. Die oben genannten Alarme finden ihre Anwendung unter anderem im Kostenverlaufsplan. Der Kostenverlaufsplan gehört zum Projektmanagement, welcher Kosten eines Projekts phasenweise oder kumuliert bereitstellt⁷¹. Im Anhang I befindet sich die Vorlage für die Erstellung eines Fakturierungsalarms in JSON und YAML Format.

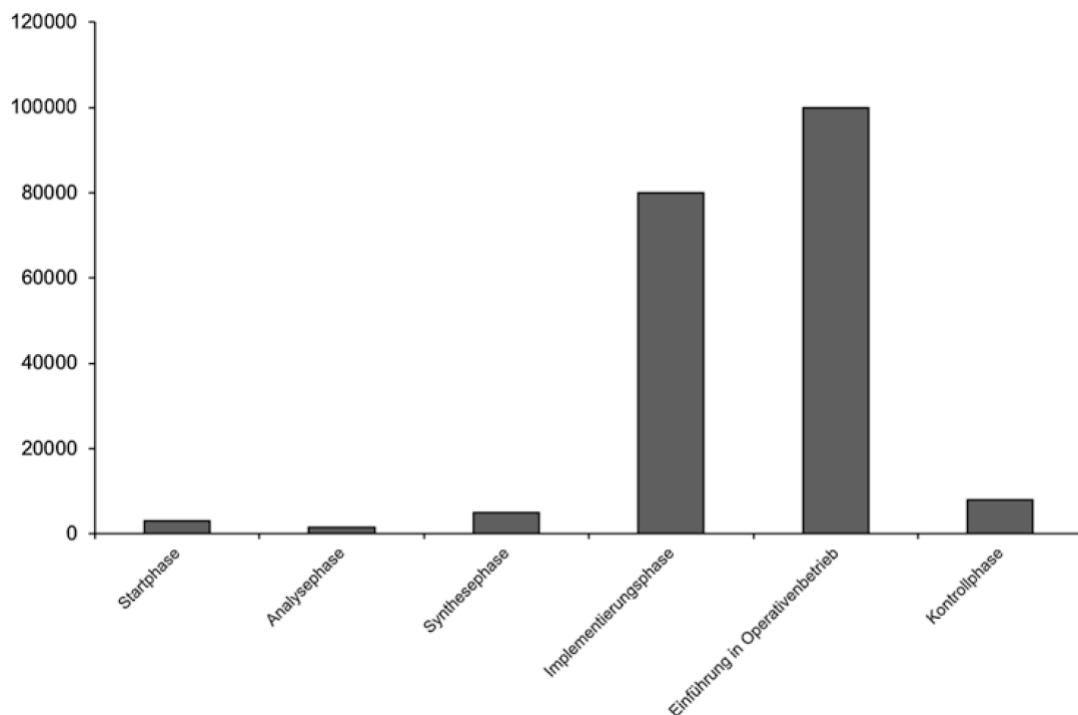


Abbildung 10
Kosten nach Projektphasen.
Kompakte Einführung in das Projektmanagement. S.97[6].

⁷¹Vgl. Kompakte Einführung in das Projektmanagement. S.96[6].

Die Abbildung 10 zeigt ein in Phasen aufgeteiltes Projekt. Für jede Phase wäre es von Vorteil, Alarme für Prozentsätze oder Überschreitungen der Budgets der einzelnen Phasen zu definieren.

4.2 AWS Cost-Explorer

Cost-Explorer erstellt Berichte über die Kosten und die Nutzung von AWS-Diensten. Darüber hinaus wird eine Kostenprognose für die nächsten Monate erstellt, welche auf die Kosten der vergangenen Monaten basiert. Die Nutzung des Cost-Explorers ist kostenlos, nur API-Aufrufe sind kostenpflichtig ⁷².

Standardberichte

Standardberichte sind vorgefertigte Berichte, die die Nutzung oder die Kosten nach einer selbstdefinierten Zeitraum zeigen. Diese zeigen eine grafische Darstellung der stündlichen, täglichen oder monatliche Kosten nach Dienst, die Abdeckung und die Auslastung von reservierten Instanzen oder die in Saving Plans Zahlungsmodell und die Ausgaben auf dem AWS Marketplace⁷³. Die Berichte über die Abdeckung und Auslastung der reservierten Instanzen wurde im TrueCar-Anwendungsfall verwendet. Dies findet sich in Unterkapitel 3.5.

Anwendungsbeispiel für Standardberichte

Eine weitere Verwendung dieser Informationen findet sich im Bereich der Marketing. Als Beispiel, ein Unternehmen, das ein Freemium-Dienst⁷⁴ anbietet. Die Marketingabteilung möchte eine Werbekampagne durchführen. Durch eine Werbekampagne werden in der Regel neue Nutzer generiert, und zwar sowohl zahlende als auch nicht zahlende Nutzer. Normalerweise gilt: Je mehr Nutzer, desto größer die Belastung für die IT-Infrastruktur. Um die im Zusammenhang mit der Werbekampagne durch neue Nutzer entstehenden Kosten zu messen, werden die tatsächlichen Kundenakquisitionskosten (CAC)⁷⁵ berechnet, wobei nur die Kosten der nicht zahlenden Nutzer berücksichtigt werden. Zur Unterscheidung

⁷²Vgl. AWS Cost Management Pricing[24].

⁷³Vgl. AWS Marketplace ist ein Einkaufskatalog für Software von Drittanbietern[36].

⁷⁴Vgl. o.V.o.J. Ein Freemium-Dienst bietet in der Regel zwei Versionen an, eine kostenlose und eine kostenpflichtige[54].

⁷⁵Vgl. Kundenakquisitionskosten sind alle anfallenden Kosten in der Customer Acquisition-Phase für ein Unternehmen[53].

zwischen alten(vor der Werbekampagne) und neuen Nutzern wird das Datum der Erstellung des Nutzerkontos verwendet. Kunden, die aufgrund der Werbekampagne von der kostenlosen zur kostenpflichtigen Version des Dienstes gewechselt haben, werden in einer anderen Kategorie⁷⁶ ausgeschlossen.

Die Formel für die Berechnung der Kundenakquisitionskosten lautet wie folgt:

Anfallende Marketingkosten (MK) addiert mit den Vertriebskosten (VK) durch die Anzahl der gewonnenen Kunden (GK).

Kosten von Nutzern, die den Dienst kostenlos in Anspruch nehmen, würden in diesem Fall in den Vertriebskosten enthalten sein. Auf diese Weise ist die Marketingabteilung in der Lage, die tatsächlichen Kosten pro zahlenden Neukunden zu berechnen, die durch die Werbekampagne generiert wurden.

Leistungskennzahlen (KPI)[Rev]

Cost-Explorer-Berichte enthalten Daten, die die Merkmale guter Leistungskennzahlen(Key Performance Indicators KPI)⁷⁷ erfüllen. Sie sind aktuell, spezifisch und in Bezug auf die Zeit messbar.

In der Abbildung 11 werden die durchschnittlichen Kosten pro Stunde für EC2-Instanzen, den Prozentsatz der Instanzen einer bestimmten Generation und der Vorgängerversionen, die Abdeckung nach Zahlungsmodell und die Verteilung des S3-Speichers nach Speicherklassen berechnet. In diesem Dashboard werden Metriken aus CloudWatch und Cost-Explorer-Berichten zusammengestellt.

Wie in der Abbildung 12 zu sehen, ist es möglich, mathematische Operationen mit den Metriken durchzuführen, um auf dem Dashboard nur die aussagekräftigsten Kennzahlen anzuzeigen.

⁷⁶Vgl. Cost-per-Action (CPA)[55].

⁷⁷Vgl. Marc Optiz. (2019). Anforderungen an Kennzahlen. Prozessorientiertes Reporting. (S.130-132). Schäffer-Poeschel Verlag für Wirtschaft.[7]

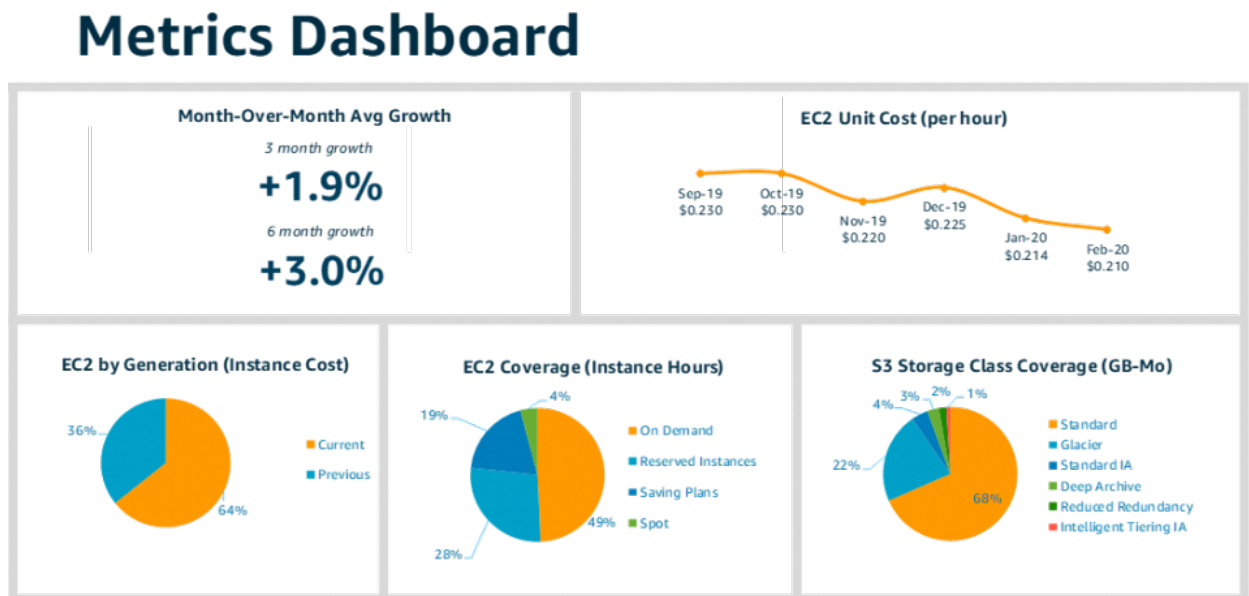


Abbildung 11
Dashboard mit Kennzahlen über EC2-Instanzen und S3-Speichereinheiten ⁷⁸

Browse	Abfragen	Grafisch dargestellte Metriken (4)	Optionen	Quelle	Add math ▲	Add query ▼
Dynamische Beschriftung hinzufügen ▼ Info				Statistik: Durchs	Prozentsatz	
<input checked="" type="checkbox"/>	Id ⓘ	Bezeichnung	Details	Statist	Summe	
<input checked="" type="checkbox"/>	q1 ⓘ	Query1 ⓘ	SELECT AVG(CPUUtilization) F...		Füllen	
<input checked="" type="checkbox"/>	e1 ⓘ	Ausdruck1 ⓘ	100*(m1/m3) ⓘ		Bytes in MB	
<input checked="" type="checkbox"/>	m1 ⓘ	GetAccountPublic...	Nutzung • CallCount • Type: API • Resc	Durchs	Bytes in GB	
<input checked="" type="checkbox"/>	m3 ⓘ	DescribeAlarmHis...	Nutzung • CallCount • Type: API • Resc	Durchschnitt ▼	Sekunde in Minute	
				5 Minuten		

Alle Funktionen

Häufig

Suchen

Sortieren

Filtern

Bedingt

Anomalie-Erkennung

SQL Query

Mit leerem Ausdruck beginnen

Abbildung 12
Mathematische Operationen an Cloud-Diensten in CloudWatch.
Quelle: CloudWatch AWS-Console

Budgetplanung

Die Budgetplanung ist eine Methode der Kostenkontrolle, die beim Start eines neuen Projekts eingesetzt wird⁷⁹. Der Cost-Explorer Berichte über die in den letzten zwölf Monaten entstandenen Kosten zusammen mit der Prognose der Kosten der kommenden zwölf Monaten tragen zu einer guten Budgetplanung bei. Durch die Möglichkeit, die in den letzten Monaten angefallenen Kosten nach bestimmten AWS-Diensten, Projekt oder Abteilung

⁷⁹Vgl. Indeed Editorial Team: Planning the budget properly. Cost Control Methods: Definitions and Examples, 2021, o.S. [45].

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 detailreicher zu planen. Welche Mittel werden wofür verwendet und welche kurz- und mittelfristigen Ziele sollen durch diesen Mitteleinsatz erreicht werden“[44]. In dem Fall dieser Arbeit sind die obengenannten Ressourcen die AWS-Dienste.

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

4.3 AWS Trusted Advisor[Rev]

AWS Trusted Advisor ist ein Werkzeug, das Empfehlungen zur Kostenreduzierung, Verbesserung der Systemverfügbarkeit und Erhöhung der Systemsicherheit gibt. Die Empfehlungen basieren auf Best-Practices, die im Laufe der Jahre durch die Betreuung von AWS-Kunden gesammelt wurden und Prüfungen, die auf dem bestehenden AWS-Konto durchgeführt wurden. In dieser Arbeit werden Empfehlungen in Bezug auf Servicekontingente und Kostenoptimierung insbesondere betrachtet, weil es sich um Empfehlungen handelt, die mit Kostenüberwachung und -optimierung zusammenhängen. Der Status von Prüfungen von Trusted Advisor sind über CloudWatch Events zugänglich.

Es ist zu berücksichtigen, dass nur limitierte Sicherheitsprüfungen (6 Prüfungen Stand November 2021) für Konten in den Plänen Developer und Basic Support kostenlos sind. Prüfungen für die Kategorie Servicekontingente zusätzlich sind kostenlos. Detaillierte Informationen und Empfehlungen von der Kategorien Kostenoptimierung, Performance und Fehlertoleranz sind nur zugänglich, wenn ein Business- oder Enterprise-Konto vorliegt⁸⁰.

Die Abbildung 13 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.

⁸⁰Vgl. AWS: Trusted Advisor o.J. o.S.[22]



Abbildung 13
 AWS Trusted Advisor Kategorien[22]

Diese Empfehlungen scheinen ein angemessener Startpunkt für die Untersuchung von AWS-Diensten zu sein. Eine genauere Untersuchung erfolgt mithilfe anderer Werkzeuge wie CloudWatch oder Cost-Explorer. Die Empfehlungen für die Kategorien Kostenoptimierung und Servicekontingente werden in der AWS-Dokumentation⁸¹ nur kurz beschrieben und sind in einem Basiskonto nicht zugänglich. Diese Kategorien lassen sich daher in eingeschränkter Weise unter den aktuellen Umständen untersuchen. Es bestehen Empfehlungen für verschiedene AWS-Dienste unter anderem für EBS, Route 53, RDS und AWS Lambda. Im Folgenden werden Empfehlungen zu EC2-Instanzen gegeben, da dies der Fokus dieser Arbeit entspricht. Empfehlungen für Amazon S3-Speichereinheiten sind im Trusted Advisor nicht verfügbar.

Empfehlungen zur Kostenoptimierung

Sollten EC2-Instanzen mit geringer Auslastung gefunden werden, wird es diese bei Trusted Advisor signalisiert. Denn diese Instanzen verursachen Kosten, welche durch die Terminierung oder das Pausieren vermieden werden können. Eine geringe Auslastung wird von AWS definiert, wenn Instanzen in den letzten 14 Tagen eine CPU-Auslastung von 10% oder weniger hatten und wenn der Netzwerkverkehr in den letzten 4 Tagen gleich oder kleiner als 5 MB war.

Buchungen von *reservierten Instanzen*, die in der letzten 30 Tage abgelaufen sind oder in den kommenden 30 Tage ablaufen werden, werden hervorgehoben. Auf diese Weise kann vermieden werden, dass die Buchung von Instanzen vergessen wird oder dass sie erneuert werden müssen, wenn sie bereits abgelaufen sind.

⁸¹Vgl. AWS Support - Benutzerhandbuch. S.59-65 und S.83-94 [38]

Empfehlungen des Cost-Explorers zu Saving Plans werden auch im Trusted Advisor angezeigt. Saving Plans sind eine mögliche Sparalternative zu reservierten Instanzen. AWS weist darauf hin, dass nur eine der beiden Maßnahmen zur Instanzreservierung durchgeführt werden sollte.

Trusted Advisor erstellt Simulationen möglicher Kombinationen von reservierten Instanzen und On-Demand-Instanzen. Dies sollte dazu dienen, die Auswahl reservierter Instanzen auf der Grundlage von AWS-Simulationen zu erleichtern.

Empfehlungen zur Servicekontingente[Rev]

In der Kategorie Servicekontingente(auch als Kontingente bekannt) werden Empfehlungen zur Vermeidung von Grenzwertüberschreitungen hervorgehoben. Sich dieser Grenzen bewusst zu sein, sollte die Möglichkeit, rechtzeitig zu handeln und es trägt zu Kostenkontrolle über die AWS-Cloud-Dienste bei.

Für Auto-Scaling-Gruppen wird es geprüft, ob deren Nutzung mehr als 80% des Kontingents beträgt. Aufgrund fehlender Informationen in der AWS-Dokumentation wird angenommen, dass eine Auto-Scaling-Gruppe als eine einzelne Recheneinheit betrachtet wird und eine Auslastung von mehr als 80% als Näherung an die Grenze der Rechenkapazität angesehen wird. Dies wird eine Anpassung der Startkonfiguration für eine bessere Skalierung zur Folge haben.

Prüfungen, die die Nutzung eines Kontingents über 80% betragen, werden auch für *On-Demand-Instanzen*, *reservierte Instanzen*, *EC2-Classic Elastic IP Addresses* und *EC2-VPC Elastic IP Addresses* angezeigt.

Trusted-Advisor Kostenerwägungen

Bei der Erwägung von Trusted-Advisor ist zu berücksichtigen, ob es kosteneffizient ist, für Support-Pläne zu zahlen. Da diese den Zugang zu allen Empfehlungen des Trusted Advisors ermöglichen. Eines der Ziele dieser Arbeit ist es, die Entstehung der Kosten auf eine praktikable Weise zu verstehen (Kostenüberwachung). Einschränkend lässt sich sagen, ob alle Empfehlungen von Trusted Advisor zu echten Einsparungen führen. Es wäre nicht sinnvoll, Kosten für AWS-Dienste wie Business- oder Enterprise Support zu übernehmen, wenn diese die möglichen Einsparungen übersteigen. Die Vorteile von Business- oder Enterprise Support-Plänen beschränken sich nicht auf Kosteneinsparungen und Kos-

tenbegrenzung, sondern tragen auch zur Sicherheit und Leistung bei. Dabei stellt sich die Frage, ob die Empfehlungen aller fünf Kategorien für die aktuelle Situation des Unternehmens benötigt werden. [[Rev]Hier die Handlungen?]

Die Preise für einen Business Support-Plan sind wie folgt definiert.

Zwischen 0 USD und 10,000 USD: 10% oder 100 USD. Je nachdem, was größer ist.

Zwischen 10,000 USD und 80,000 USD: 7%.

Zwischen 80,000 USD und 250,000 USD: 5%.

Ab 250,000 USD: 3%⁸².

Die Preise für einen Enterprise Support-Plan sind wie folgt definiert.

Zwischen 0 USD und 150,000 USD: 10% oder 15,000 USD. Je nachdem, was größer ist.

Zwischen 150,000 USD und 500,000 USD: 7%.

Zwischen 500,000 USD und 1,000,000 USD: 5%.

Ab 1,000,000 USD: 3%⁸³.

Die Prozentsätze basieren auf der monatlichen Gebühr für AWS-Dienste.

Beide Pläne bieten rund um die Uhr technischen Support durch AWS-Ingenieure und andere zusätzliche Dienstleistungen, auf die in dieser Arbeit nicht weiter eingegangen wird. Die Preise der Support-Pläne geben einen Hinweis darauf, ob die zu zahlende Empfehlungen von Trusted Advisor zur Kostenoptimierung und -überwachung kosteneffizient würden.

⁸²Vgl. AWS Support Plan Pricing - Business Support-Plan, 2021, o.S. [39]

⁸³Vgl. AWS Support Plan Pricing - Business Enterprise-Plan, 2021, o.S. [39]

4.4 Überwachungswerkzeuge gemäß ihrer Verwendung

[[Rev]NOCH NICHT VOLLSTÄNDIG] Abbildung 14 fasst die Überwachungswerkzeuge zusammen und listet deren Einsatzmöglichkeiten auf.

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 14
Überwachungswerkzeuge gemäß ihrer Verwendung
Eigene Darstellung[13, 22, 23].

Fazit

In diesem Kapitel wurde gezeigt, dass es mit CloudWatch möglich ist, Alarme auf Basis von Ereignissen einzurichten, die mit Amazon SNS oder externen E-Mail-Adressen kommunizieren. 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, welcher 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.

5 Optimierungsmaßnahmen

[Rev]Die mit den Überwachungswerkzeuge gesammelte Informationen, bilden die Grundlage für die Optimierungsmaßnahmen. Sollen sich In diesem Kapitel werden die mithilfe der Werkzeuge gewonnenen Informationen genutzt, um über die am besten geeigneten Optimierungsmaßnahmen zu entscheiden.

5.1 EC2 Auto Scaling

Auto Scaling oder automatische Skalierung von Instanzen ist es hilfreich, um die richtige Anzahl von EC2-Instanzen zur Verfügung zu haben, um die Anwendungslast dynamisch abzudecken⁸⁴. Dieses wird als *horizontale Skalierung* bezeichnet⁸⁵.

Die Abbildung 15 zeigt das wechselnde Verhalten einer Beispielanwendung, die vor allem unter der Woche genutzt wird. Am Wochenende sinkt die Nachfrage nach Rechnerkapazität auf weniger als 25 % und lässt den Rest der Kapazität ungenutzt.

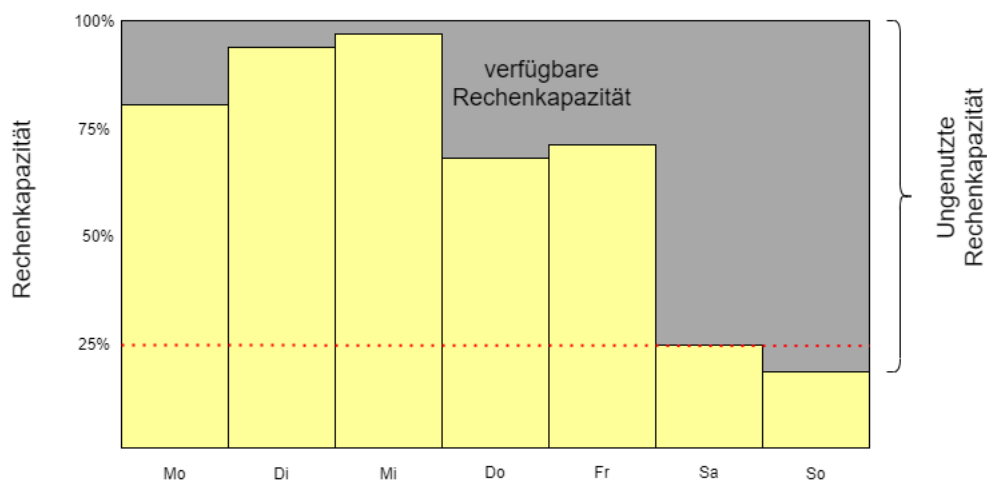


Abbildung 15
Ungenutzte Rechenkapazität ohne automatische Skalierung.
Quelle: Eigene Darstellung mit fiktiven Angaben.

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

⁸⁴Vgl. Was ist Amazon EC2 Auto Scaling? S.9[33]

⁸⁵Vgl. Die Grundbedeutung der horizontalen Skalierung ist, dass Systeme durch zusätzliche Komponenten erweitert werden. Im Gegensatz dazu bedeutet der Begriff "vertikale Skalierung", dass einer einzelnen Komponente zusätzliche Leistungsfähigkeiten und Ressourcen hinzugefügt werden. o.S.[63]

Auto Scaling Group

Die Instanzen, die zur Deckung der erforderlichen Rechenkapazität zur Verfügung stehen, werden in einer *Auto-Scaling-Gruppe* (*Auto Scaling Group*) zusammengestellt. Diese Gruppe von Instanzen wird in AWS als Auto-Scaling-Gruppe bezeichnet. Bei der Erstellung einer Auto-Scaling-Gruppe wird eine minimale, gewünschte und maximale Anzahl von Instanzen definiert.

Die Abbildung 16 zeigt die gewünschte Instanzen einer Auto-Scaling-Gruppe, welche beim Start der Auto-Scaling-Gruppe gestartet werden. Die minimale und maximale Anzahl von Instanzen sind die Grenzwerte für die Auto-Scaling-Gruppe.

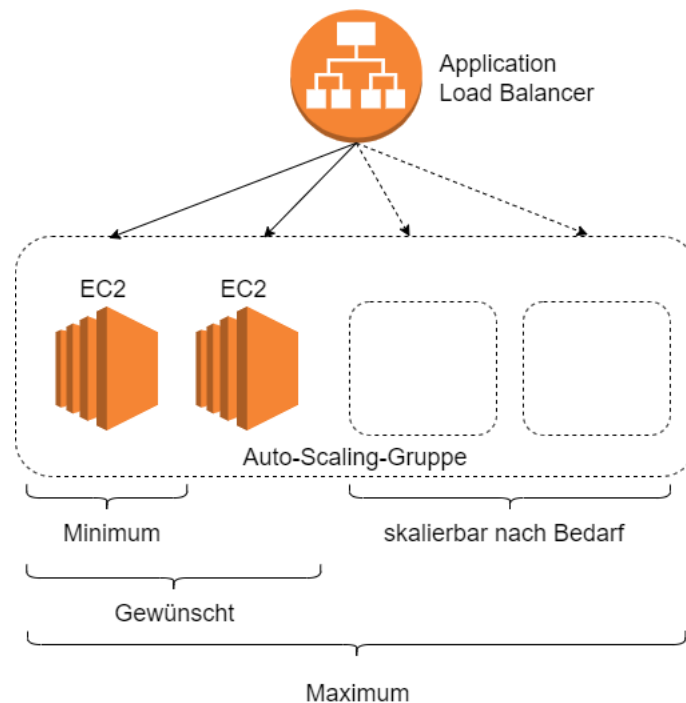


Abbildung 16

Auto-Scaling-Gruppe nach den Anzahl der Instanzen und die Umleitung der Datenverkehr durch dem Application Load Balancer.

Quelle: Eigene Darstellung basiert auf Amazon EC2 Auto Scaling - Benutzerhandbuch. S.9[33].

Elastic Load Balancing[Rev]

Ein *Elastic Load Balancer* ist für die Verwaltung eingehender Anfragen zuständig, indem es den Datenverkehr auf alle laufenden EC2-Instanzen umleitet⁸⁶. Dies sorgt dafür, Instanzen mit einem ausgeglichener CPU-Auslastung arbeiten. Die Abbildung 16 zeigt einen Application-Load-Balancer, welcher den Datenverkehr auf die Instanzen einer Auto-Scaling-Gruppe verteilt.

5.1.1 Zeitgesteuerte Skalierung

Nicht produktive Umgebungen

In einem On-Premise-System mache es, wenn überhaupt, einen kleinen Unterschied bei den Kosten, dass Instanzen die ganze Zeit aktiv bleiben⁸⁷. Im Gegensatz dazu ist es bei On-Demand-Zahlungsmodelle sinnvoll Zeiträume zu definieren, in denen Instanzen abgeschaltet werden können, um deren Nutzung zu reduzieren. Bei Systemen, die nur tagsüber und unter der Woche in Betrieb sein müssen, kann dies eine Einsparung von bis zu 67% bedeuten. Wenn zum Beispiel Test- und Beta-Umgebungen von Montag bis Freitag von 7 bis 20 Uhr laufen würden.

Die Abbildung 17 zeigt die Kostenberechnung einer nicht produktiven Umgebung (z.B. Test, Dev oder Beta) mit On-Demand-Instanzen. Diese Umgebung wird nur von Montag bis Freitag von 7:00 bis 20:00 Uhr genutzt. In der rechten Spalte werden die Kosten für Instanzen berechnet, wenn sie immer aktiv bleiben. In der linken Spalte wurde eine Berechnung durchgeführt, bei der die Instanzen nur dann eingeschaltet werden, wenn die Instanzen nach einem Zeitplan gesteuert würden.

Die Abbildung zeigt am Ende den Prozentsatz und den Betrag(in Euros) der möglichen Einsparungen, wenn die Instanzen nach einer Zeitplan steuert werden würden.

⁸⁶Vgl. In diesem Fall beschränkt auf den Application Load Balancer. Amazon Elastic Container Service Entwicklerhandbuch - Load Balancer-Typen - S.617[40]

⁸⁷Anders Lisdorf, 2021, S. 153[4]

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 17

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

Quelle des Stundensatzes: AWS Pricing Calculator⁸⁸

⁸⁸Der Stundensatz wurde am 23.11.2021 mit dem AWS Pricing Calculator ermittelt für Linux Instanzen in Frankfurt mit 4vCPUs, 16 GB Arbeitsspeicher und Instanz-Familie t4g.xlarge in On-Demand-Zahlungsmodell[18].

5.1.2 Dynamisches Auto Scaling[Rev]

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 18 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 passend, wenn das Ziel darin besteht, ungenutzte Cloud-Dienste abzuschalten. Als Konsequenz der Abschaltung von ungenutzten Cloud-Diensten folgt die Reduzierung von Kosten.

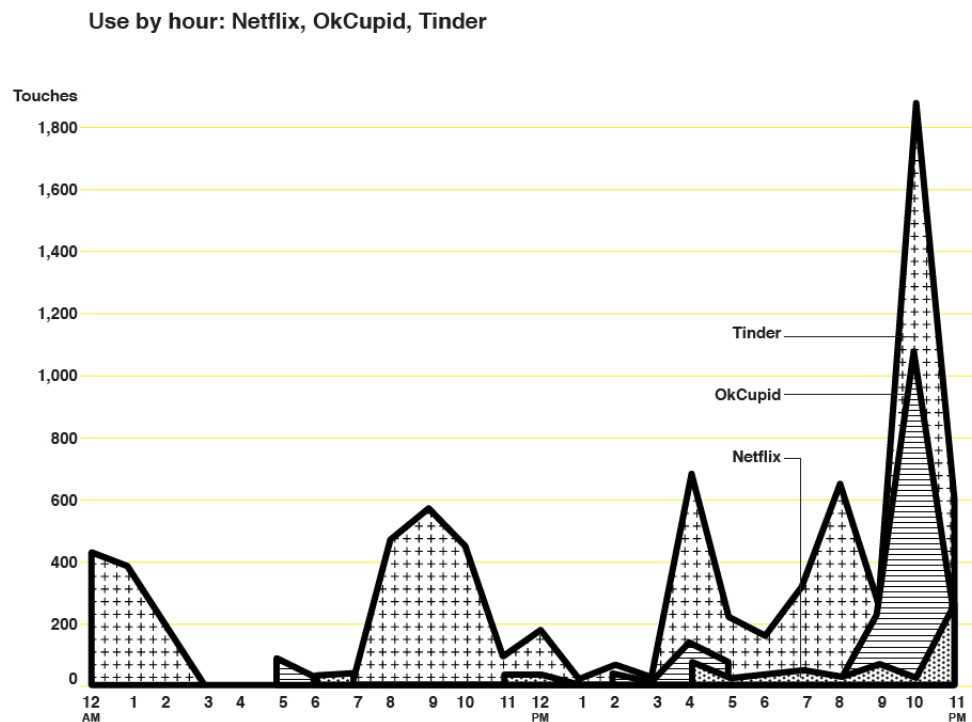


Abbildung 18

DScout's Study: "Putting a Finger on Our Phone Obsession".

Nutzung pro Stunde von Netflix, OkCupid und Tinder während des Tages[67].

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

Die für die automatische Skalierung erforderlichen Metriken wurden näher im Unterkapitel 4.1 erwähnt. Eine der Metriken, die von AWS benutzt wird, ist die gesamte CPU-Auslastung(CPU-Utilization). 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 oder weniger 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 des Nachfrageanstiegs und der Initialisierung der Instanz vermieden oder verkürzt werden. Anders als Zeitgesteuerte Skalierung ist es nicht notwendig, die Verhaltensmuster der Anwendungen zu analysieren.

5.2 S3 Optimierung

In diesem Unterkapitel werden Maßnahmen zur Speicheroptimierung für Amazon S3 beschrieben. Jedem Objekt in Amazon S3 ist eine Speicherklasse zugewiesen. Die Speicherklassen werden nach der Zugriffshäufigkeit auf die Objekte unterschieden und sind für verschiedene Szenarien konzipiert. Es gibt Speicherklassen für den häufigen und den seltenen Zugriff⁸⁹. Der Preis bei Amazon S3 wird pro GB berechnet und ist umso niedriger, je geringer der Zugriff auf die Objekte ist⁹⁰. Um die Speicherkosten zu optimieren, ist es daher notwendig, die richtige Speicherklassen für die jeweilige Applikation zu wählen, weil die Speicherkosten durch ihre Klasse berechnet werden.

⁸⁹Vgl. AWS: Amazon Simple Storage Service - User Guide. S.709.[20]

⁹⁰Vgl. AWS S3 Pricing [10]

5.2.1 Die richtige Speicherklassen wählen[Rev]

Um die richtige Wahl zu treffen, müssen die Anforderungen der Applikation verstanden werden. Ärztliche Patientenakten und *Instagram-Stories*⁹¹ sind zwei Beispiele für Daten, die nach deren Erstellung für einen Mindestzeitraum oder auf unbestimmte Zeit aufbewahrt werden. In Deutschland müssen ärztliche Patientenakten mindesten zehn Jahre aufbewahrt werden⁹². *Instagram* verwendet die von seinen Nutzern bereitgestellten Informationen, einschließlich der Metadaten von Bildern, um andere Instagram- und *Facebook*-Produkte zu empfehlen⁹³. Die Zugriffshäufigkeit und die Aufbewahrungszeit sind die zwei Hauptkriterien für die Verschiebung von Daten zwischen Speicherklassen⁹⁴.

([Rev] UMFORMULIEREN :) Objekte werden in Behältern gespeichert, die Buckets genannt werden. Daten werden über einen längeren Zeitraum gespeichert aufgrund der vorgeschriebenen Anforderungen oder weil per Gesetz auf die Informationen in der Zukunft zugegriffen werden muss. Zusätzlich, wenn auf die Daten nicht häufig zugegriffen wird, sind Glacier und Glacier Deep Archive passende Speicherklassen. Die Entscheidung für eine bestimmte Speicherkategorie ist jedoch nicht immer so leicht zu treffen. 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 verschieben.

5.2.2 Lebenszyklus-Konfiguration

Die *Lebenszyklus-Konfiguration* oder *lifecycle policy* ist eine Maßnahme zur Optimierung von Amazon S3-Speichereinheiten. Eine S3-Lebenszykluskonfiguration beschreibt in einer XML-Datei Regeln und Aktionen für die Verschiebung in unterschiedlichen Speicherklassen von Objekten. Die Verschiebung von Objekten verursachen Kosten. Ein Beispiel von diesen Kosten und mögliche Einsparungen werden in Abbildung 19 vorgestellt.

Um konkretere Regeln zu definieren, ist es möglich Tags zu verwenden und somit

⁹¹Vgl. Bei Instagram Stories handelt es sich um kurzen visuellen Content in der Regel Bilder oder kurze Videos, die nach 24 Stunden automatisch aus der Applikation Instagram verschwinden(Stand November 2021).[50]

⁹²Vgl. Nach dem Bürgerlichen Gesetzbuch (BGB) § 630f müssen Patientenakten zehn Jahren nach Abschluss der Behandlung aufbewahrt werden, soweit nicht nach anderen Vorschriften andere Aufbewahrungsfristen bestehen. [43]

⁹³Vgl. Instagram macht keine genauen Angaben darüber, wie lange die Nutzerdaten aufbewahrt werden, sondern gibt nur an, dass sie so lange wie nötig aufbewahrt werden. Hilfebereich Instagram: VII. Datenspeicherung, Deaktivierung und Löschung von Konten[51].

⁹⁴Vgl. AWS: Amazon Simple Storage Service - User Guide. S.711.[20]

eine Unterscheidung zwischen Objekten mit verschiedenen Tags zu treffen. Es ist zum Beispiel möglich, alle Objekte mit dem Tag-Wert: *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,

⁹⁵Vgl. AWS: Amazon Simple Storage Service - User Guide. S.701.[20]

5.2.3 Anwendungsbeispiel für eine Lebenszyklus-Konfiguration

Zur Veranschaulichung der Verschiebung von Objekten zwischen Speicherklassen wird der folgende Anwendungsfall vorgestellt. In diesem Fall wird der Punkt als Dezimaltrennzeichen und das Komma als Tausendertrennzeichen verwendet.

Ein Sicherheitsunternehmen muss Sicherheitsvideos speichern, die aktuell im Summe 120 TB groß sind. Viele von ihnen werden mindestens 5 Jahre lang aufbewahrt, falls sie vor Gericht als Beweismittel dienen sollten. 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. Jedes Video 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 zweites Szenario betrachtet, in dem die Videos je nach Alter verschoben werden. Im zweiten Szenario müssen die Kosten für die Verschiebung zwischen Speicherklassen berücksichtigt werden. Die Verschiebung erfolgt durch eine Lebenszyklus-Konfiguration wie in der Unterkapitel 5.2.2 beschrieben. Zum besseren Verständnis wird angenommen, dass 20% der Dateien in S3 Standard Infrequent Access und 30% in S3 Glacier gespeichert werden.

Durchschnittliche Dateigröße	3,4	GB
Anzahl der Dateien	36,141	Überwachungsvideos
Gesamtspeicher	122,880	GB
	120	TB

Ausschließlich S3-Standard verwenden		
	S3 Standard (erste 51200GB)	S3 Standard (Nächste 450 TB)
Speicherplatz in GB	51,200	71,680
Preis pro GB	\$0.0245	\$0.0235
Speicherverteilung	42%	58%
Anzahl der Dateien	15,059	21,082
Übertragungsgebühr (pro 1.000 Aufrufe)	-	-
Kosten für Verschiebung	0	0
Speicherkosten	\$1,254.40	\$1,684.48
Monatliche 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	51,200	10,240	24,576	36,864
Preis pro GB	\$0.0245	\$0.0235	\$0.0136	\$0.0045
Speicherverteilung	42%	8%	20%	30%
Anzahl der Dateien	15,059	3,012	7,228	10,842
Übertragungsgebühr (\$0.01/1,000 Aufrufe)	-	-	\$0.0100	\$0.0360
Kosten für Verschiebung	0	0	\$0.72	\$3.90
Speicherkosten	\$1,254.40	\$240.64	\$334.23	\$165.89
Monatliche Gesamtkosten	\$1,999.79			

Abbildung 19
Kostenvergleich durch Nutzung von unterschiedlichen Speicherklassen.

Quelle: Eigene Darstellung mit Stundensätze der S3-Preise⁹⁶.

Bei der Berechnung wurden die Kosten für das Verschieben von Videos zwischen Speicherklassen berücksichtigt. Anhand der Berechnungen in der Abbildung 19 lässt sich erkennen, dass ein Einsparungspotenzial von rund 1,000(Eintausend) USD pro Monat besteht, indem die notwendigen Regeln aufgestellt werden, um einen Teil der Videos in anderen Speicherklassen zu verschieben, welche niedrigere Preise bieten.

⁹⁶Vgl. AWS S3 Pricing[10]

5.2.4 Intelligent-Tiering

Intelligent-Tiering verschiebt Objekte auf der Grundlage von Zugriffsmustern. Diese Speicherklasse ist ideal für Objekte mit wechselnden oder unbekannten Zugriffsmustern. Wie die Senior Product Managerin für S3 Ruhi Dang erklärt, haben einige Unternehmen weder die Zeit noch das Geld, um 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 USD für Speicher ausgeben ⁹⁷. Die Abbildung 20 zeigt, wie Objekte in Abhängigkeit davon, ob auf sie zugegriffen wurde oder nicht, verschoben werden.

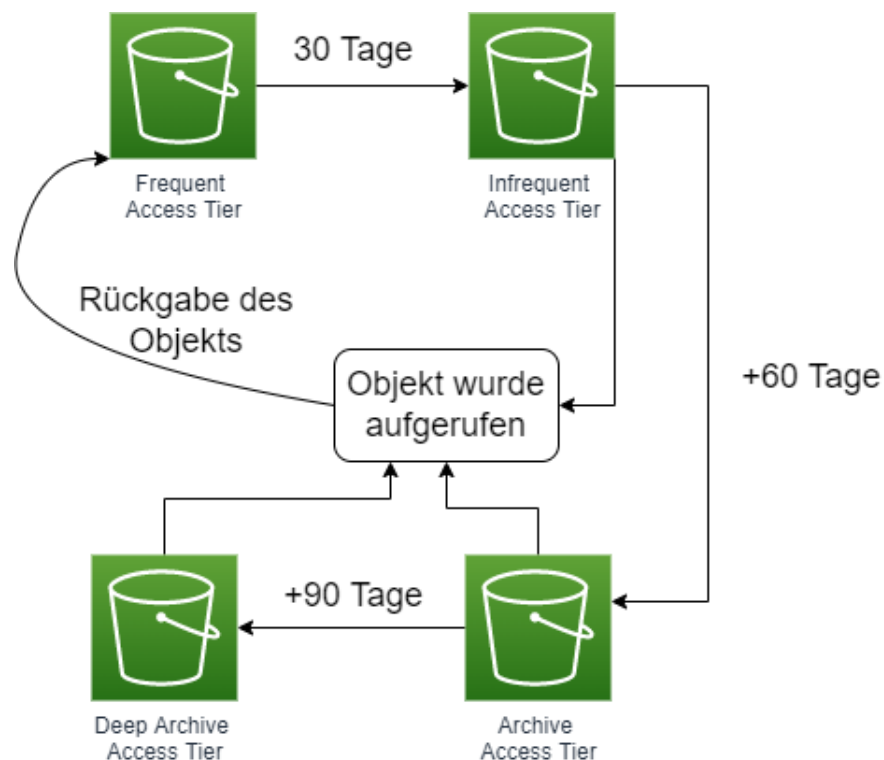


Abbildung 20
Funktionsweise von Intelligent-Tiering

Quelle: Eigene Darstellung auf der Grundlage von der Funktionsweise von Intelligent-Tiering⁹⁸.

Wird ein Objekt zu einem späteren Zeitpunkt aus der Ebene der seltenen Zugriffe aufgerufen, wird dieses automatisch in eine Speicherklasse der häufigen Zugriffe zurückversetzt.

⁹⁷Vgl. AWS re:Invent 2019: Guidelines and design patterns for optimizing cost in Amazon S3. Minute: 21:12 [17]

⁹⁸Vgl. Amazon Simple Storage Service - User Guide. S.715[20]

Das in Unterkapitel 5.2.3 dargestellte Szenario wurde mit dem AWS Pricing-Calculator für S3⁹⁹ unter Verwendung der Speicherklasse S3 Intelligent-Tiering berechnet. Da bei der Berechnung des Unterkapitels 5.2.3 eine Speicherklasse für häufigen Zugriff, eine für seltenen Zugriff und eine für Archivierung verwendet wurde, wurden für die Berechnung mit Intelligent-Tiering vergleichbare Speicherebenen ausgewählt. In diesem Fall wurden ausschließlich die Ebenen *Frequent-Access*, *Infrequent-Access* und *Instant-Archive-Access* ausgewählt.

Die Speicherzuweisung für die 120 TB in den Speicherebenen war wie folgt:

Frequent-Access-Tier: 50%

Infrequent-Access-Tier: 20%

Instant-Archive-Access: 30%

▼ Berechnungen anzeigen

Einheitenumwandlung

S3 INT-Speicher: 120 TB pro Monat x 1024 GB in TB = 122880 GB pro Monat

Prozentsatz des Speichers in INT-Frequent-Access-Stufe: $50 / 100 = 0.5$

Prozentsatz des Speichers in Stufe INT-Infrequent-Access (% des Speichers, auf den in den letzten 30 Tagen nicht zugegriffen wurde): $20 / 100 = 0.2$

Prozentsatz des Speichers in der INT-Archive-Instant-Access-Stufe (% des Speichers, auf den in den letzten 90 Tagen nicht zugegriffen wurde): $30 / 100 = 0.3$

Von S3 Select gescannte Daten: 120 TB pro Monat x 1024 GB in TB = 122880 GB pro Monat

Preisberechnungen

0,50 Multiplikator für häufigen Zugriff x 122.880 GB = 61.440,00 GB (Gesamter Speicher für häufigen Zugriff)

Tiered price for: 61440.00 GB

51200 GB x 0.0245000000 USD = 1254.40 USD

10240 GB x 0.0235000000 USD = 240.64 USD

Gesamtstufenkosten: 1254.40 USD + 240.64 USD = 1495.0400 USD (S3-INT-Speicher, Kosten der Stufe für Frequent Access)

Abbildung 21

Berechnung für die Verwaltung von 120 TB mit AWS Pricing-Calculator für S3 Intelligent-Tiering(1).

Quelle: eigene Darstellung von AWS Pricing-Calculator [19].

⁹⁹Vgl. AWS Pricing Calculator S3[19].

In der Abbildung 22 werden die Kosten gezeigt, die durch jede Speicherebene anfallen (mit orange markiert).

Darüber hinaus werden die folgenden Kosten berechnet (mit blau markiert):

- Überwachung und Automatisierung von 36.141 Objekten.
- Leseanfragen (GET-Anfragen) von 18.070 Objekten, welche 50% der Gesamtzahl der Videos entsprechen, wie in den Anforderungen im Unterkapitel 5.2.3 definiert.
- Scannen von Objekten, die allen Videos und 120 TB entsprechen.

Kosten für Frequent Access Tier: 1.495,04 USD

0,20 Multiplikator für seltenen Zugriff x 122.880 GB (Gesamter S3-INT-Speicher) = 24.576,00 GB (Gesamtspeicher für seltenen Zugriff)

24.576,00 GB x 0,0135 USD = 331,776 USD (S3-INT-Speicher, Kosten der Stufe für Infrequent-Access)

Kosten für Infrequent Access Tier: 331,776 USD

0,30 Archivierungsmultiplikator für sofortigen Zugriff x 122.880 GB (Gesamter S3-INT-Speicher) = 36.864,00 GB (Gesamter Archivspeicherplatz für sofortigen Zugriff)

36.864,00 GB x 0,005 USD = 184,32 USD (S3 INT Storage, Kosten für die Stufe Archive Instant Access)

Kosten für die Stufe Instant Archive Access: 184,32 USD

Kosten Archive Access Tier: 0 USD

Kosten für Deep Archive Access Tier: 0 USD

36.141 Überwachung und Automatisierung (Objekte pro Monat) x 0,0000025 USD pro Objekt = 0,0904 USD (Kosten für die Überwachung und Automatisierung von Objekten)

18.070 GET-Anfragen in einem Monat x 0,00000043 USD pro Anfrage = 0,0078 USD (Kosten für S3-INT-GET-Anfragen)

122.880 GB x 0,00225 USD = 276,48 USD (Kosten für gescannte S3-INT-Daten)

1.495,04 USD + 331,776 USD + 184,32 USD + 0,0904 USD + 0,0078 USD + 276,48 USD = 2.287,71 USD (Gesamtkosten für S3-INT-Speicher, Anforderungen, Auswahl, gescannte und Abrufkosten)

Kosten für S3 Intelligent-Tiering (S3 INT) (monatlich): 2,287.71 USD

Abbildung 22

Berechnung für die Verwaltung von 120 TB mit
AWS Pricing-Calculator für S3 Intelligent-Tiering(2).
Quelle: eigene Darstellung mit AWS Pricing-Calculator [19].

Unterscheidung zwischen Intelligent-Tiering und Lebenszyklus-Konfiguration[Rev]

Ein Unterschied zwischen der Verwendung von Intelligent-Tiering und einer Lebenszyklus-Konfiguration besteht darin, dass bei Intelligent-Tiering die Objekte in Ebenen von seltenen Zugriff automatisch auf eine Ebene für häufigen Zugriff zurückgegeben werden. Siehe Abbildung 20.

Ein weiterer Unterschied ist, dass bei einer Lebenszyklus-Konfiguration die Tage der Verschiebung zwischen den Speicherklassen und die Speicherklassen leicht verändert werden können. Dies ist standardmäßig bei Intelligent-Tiering bereits festgelegt.

Es gibt jedoch die Möglichkeit, Objekte aus Intelligent-Tiering in andere Speicherklassen zu verschieben (mit Einschränkungen¹⁰⁰), indem man eine Lebenszyklus-Konfiguration verwendet. Dies gäbe die Möglichkeit, Objekte für einen längeren Zeitraum in einer bestimmten Speicherklasse aufzubewahren und Richtlinien von Intelligent-Tiering zu überspringen.

¹⁰⁰Vgl. AWS: Amazon Simple Storage Service - User Guide. S.724.[20]

Zusammenfassung und Ausblick

[Rev]

Kurzdarstellung der Inhalte

Im Kapitel 2 wurde die Bedeutung von Cloud-basierten Systemen wurde bestätigt und verstärkt. Zum einen durch die Statistiken[68, 69, 71] über die Nutzung von Cloud-Systemen weltweit. Zum anderen durch die Anzahl von Unternehmen, auch in Deutschland[2], mit erfolgreicher Implementierung von Public Cloud-basierten Systemen. Es hat sich gezeigt, dass viele Unternehmen derzeit Schwierigkeiten haben, auf die Cloud umzusteigen, weil ihnen die technische Qualifikation fehlt[Zitat].

Kapitel 3 Weil EC2 den Großteil der Kosten von Cloud-Diensten ausmacht, wurden deren Zahlungsmodelle untersucht und vorgestellt wann welches Zahlungsmodell das geeignetste ist. +Anhand einer Berechnung von On-Demand Instanzen mit zeitgesteuerter Skalierung, wurden die möglichen Einsparungen ermittelt. +Es hat sich gezeigt, dass die Kombination von Zahlungsmodellen sich zur Kostensenkung anbietet. Dies unter Einbeziehung von EC2-Flotten für die erwartete und On-Demand Instanzen für die variable Nutzung.

Wie in dem Anwendungsfall von Truecar Inc. in Unterkapitel 3.5 wurde bestätigt, dass die korrekte Berechnung der künftig nötigen reservierten Instanzen und deren spätere Überwachung, erhebliche Einsparungen erzielen.

Im Kapitel ?? wurden drei Überwachungswerkzeuge untersucht mit denen Optimierungsmaßnahmen zu ergreifen sind. Die damit gesammelte Informationen dienen in Bereichen der Betriebswirtschaft und sollte eine bessere Kontrolle über die Kosten von Cloud-Diensten ermöglichen.

Cost-Explorer hat gezeigt wie mit Berichten einen umfangreichen Überblick der Nutzung und Kosten zu verschaffen ist. Die zwei vorgestellten Einsatzmöglichkeiten sind die operative Budgetplanung und Verfolgung von Lesitungskennzahlen(KPIs). Die Metriken bei CloudWatch haben gezeigt das Potenzial zu haben, die Verfolgung von Lesitungskennzahlen zu ergänzen. Darüber hinaus mit den Dashboards und Benachrichtigungen haben gezeigt, dass wie sie in echter Zeit über den Stand der Cloud-Diensten informieren. Trusted Advisor wurde aufgrund des fehlenden Zugangs zu einem Business- oder Enterprise-Plan nur eingeschränkt untersucht. Auch der Dokumentation in Bezug auf Trusted Advisor was nicht ausreichend, um Berechnungen, Prüfungen oder Ähnliches durchführen zu kön-

nen. Die Informationen, die gesammelt werden konnten, geben jedoch Hinweise auf die in dieser Arbeit behandelten Optimierungsmaßnahmen. Zum Beispiel die Verwendung von reservierten Instanzen und das Abschalten von Instanzen mit geringer Auslastung.

Im Kapitel 5 wurde anhand von definierten Arbeitszeiten mögliche Einsparungen für eine nicht-produktive Entwicklungsumgebung mit On-Demand EC2-Instanzen berechnet. Außerdem wurde erklärt, wie Auto Scaling-Gruppen, Load Balancer und Auto Scaling die Leistung einer balancierten und vertikal skalierbaren Serverinfrastruktur ermöglichen können.

Mit Fokus auf Amazon S3 wurden Speicherklassen und das Verschieben von Objekten zwischen ihnen mithilfe von Lebenszyklus-Konfiguration und Intelligent-Tiering vorgestellt.

Kurzdarstellung Problem-Lösungsweg-Ergebnisse

Problem: Kosten von Cloud-Diensten sind für ihre Verbraucher(Unternehmen) nicht verständlich. Als Teil des Problems gehört die fehlende Fachkraft.

Lösungsweg: Schulungen+(Überwachungswerkzeuge+Optimierungsmaßnahmen)

Ergebnisse: bessere Kostenübersicht und -kontrolle, durch die Nutzung von den passenden Werkzeugen/Diensten.

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

Werkzeuge und Maßnahmen sind da und warten darauf genutzt zu werden...?

Kritische Bewertung (sofern nicht bereits im Hauptteil geschehen)

Nur mit AWS-Diensten zu arbeiten könnte zu (Abhängigkeit von dem Cloud-Anbieter führen; dafür gibt es ein Fachbegriff), deswegen lädt Multi-Cloud ein zu einzeln

Offene Probleme/Themen

AWS-Organizations, IAM, Mit Serverless können Kosten optimiert werden, möglicherweise würde die Komplexität der Anwendung zunehmen., AWS Cost Anomalies, CloudFormation? +Werkzeuge Testen, Maßnahmen ergreifen und Ergebnisse messen.

Richtung der zukünftigen/möglichen Arbeiten - Weitere Forschungen

Während der Entwicklung dieser Arbeit wurden S3, Spot-Instanzen, Cost-Explorer, Cloud-Watch und Trusted Advisor (mit Einschränkungen) mit dem kostenlosen AWS-Kontingent getestet. Es bleibt offen, die Überwachungswerkzeuge zu verwenden und die Optimierungsmaßnahmen in einer echten IT-Infrastruktur zu ergreifen. Zum Beispiel bei Rechenlasten(Wort)?, die nicht vorhersehbar sind, oder bei Entwicklungsumgebungen, die nach Arbeitszeiten ein- und ausgeschaltet werden können. Für die Datenspeicherung mit unterschiedlichen Zugriffsmustern werden S3 intelligent-Tiering oder Policies zu testen.

Erläuterung, warum welche Aspekte in der Arbeit nicht erläutert

Es wurden zwei AWS-Dienste untersucht und nicht mehr, um den Fokus auf die meist genutzte und representative Cloud-Dienste von AWS zu halten. Aus diesem Grund wurden Optimierungsmaßnahmen von Trusted Advisor für Datenbanken, Netzwerk, AWS-Lambda und mehr nicht berücksichtigt. ANDERE ASPEKTE(Wenn Zeit gibt):

Bewusstsein in der gesamten Organisation entwickeln

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 IT-Manager, jeder sollte wissen, dass es so einfach ist, Cloud-Dienste mit ein paar Klicks zu beauftragen¹⁰¹. Diese können in kurzer Zeit ungewünschte Kosten verursachen oder sogar über Jahre hinweg wirtschaftliche Schäden verursachen.

Die richtige Personen finden, Ownership/Commitment 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

¹⁰¹Vgl. Plusserver: Kostenoptimierung in AWS, S.5[61].

Ziele zu erreichen.

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.

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 AWS-Diensten 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.

Quellenverzeichnis

Literatur

- [1] AWS Certified Solutions Architect - Associate (SAA-C02)

ISBN: 9780137325160

(Abgerufen am 02.11.2021)

- [2] Marceil Schweitzer und Ernst Troßmann. Break-even-Analysen Methodik und Einsatz.

https://www.wiso-net.de/document/DUHU__9783428490882522

ISBN: PDF 978-3-428-49088-2

(Erscheinungsjahr 1998)

- [3] V.Bach, & H. Österle. Business Knowledge Management: Wertschöpfung durch Wissensportale.

ISBN: 3-540-42804-6

- [4] Anders Lisdorf (2021): *Cloud Computing Basics: a Non.-Technical Introduction*. Apress.

ISBN-13 (pbk): 978-1-4842-6920-6

- [5] Helmut Krcmar (2015): Informationsmanagement. 6. Auflage.

ISBN: 978-3-662-45863-1 (eBook)

- [6] Kompakte Einführung in das Projektmanagement. Theo PetersNicole Schelter

<https://link.springer.com/book/10.1007%2F978-3-658-31194-0>

ISBN: 978-3-658-31194-0

- [7] Prozessorientiertes Reporting

<https://content-select.com/de/portal/media/view/5e419784-8730-4de1-a69d-561eb0dd2d03?forceauth=1>

(Abgerufen am 11.12.2021)

ISBN: 9783791046556

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#:~: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 EC2
<https://calculator.aws/#/createCalculator/EC2> (Abgerufen am 23.11.2021)
- [19] AWS Pricing Calculator S3
<https://calculator.aws/#/createCalculator/S3> (Abgerufen am 13.12.2021)
- [20] 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)
- [21] 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)
- [22] AWS Trusted Advisor
<https://aws.amazon.com/de/premiumsupport/technology/trusted-advisor/> (Abgerufen am 26.11.2021)
- [23] AWS Cost Explorer
<https://aws.amazon.com/de/aws-cost-management/aws-cost-explorer/>
(Abgerufen am 26.11.2021)
- [24] AWS Cost Management Pricing
<https://aws.amazon.com/de/aws-cost-management/pricing/>
(Abgerufen am 30.11.2021)
- [25] 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)
- [26] 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)
- [27] AWS EC2 Spot Instanzen-Anfragen und Preisverlauf
<https://console.aws.amazon.com/ec2sp/v1/spot/home?>
(Abgerufen am 01.12.2021)
- [28] 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)
- [29] AWS X-Ray Developer Guide: What is AWS X-Ray?
<https://docs.aws.amazon.com/xray/latest/devguide/xray-guide.pdf#aws-xray>
(Abgerufen am 03.12.2021)

-
- [30] AWS CloudTrail User Guide Version 1.0: What Is AWS CloudTrail?
<https://docs.aws.amazon.com/awscloudtrail/latest/userguide/awscloudtrail-ug.pdf#cloudtrail-user-guide>
(Abgerufen am 03.12.2021)
- [31] AWS – Allgemeine Referenz - Referenzhandbuch
https://docs.aws.amazon.com/general/latest/gr/aws-general.pdf#aws_tagging (Abgerufen am 04.12.2021)
- [32] AWS – Amazon SNS
<https://aws.amazon.com/de/sns/> (Abgerufen am 04.12.2021)
- [33] Amazon EC2 Auto Scaling - Benutzerhandbuch
https://docs.aws.amazon.com/de_de/autoscaling/ec2/userguide/as-dg.pdf#what-is-amazon-ec2-auto-scaling
(Abgerufen am 05.12.2021)
- [34] AWS CloudFormation - Benutzerhandbuch
https://docs.aws.amazon.com/de_de/AWSCloudFormation/latest/UserGuide/cfn-ug.pdf#quickref-cloudwatch
(Abgerufen am 05.12.2021)
- [35] AWS Single Sign-On
https://aws.amazon.com/single-sign-on/?nc1=h_ls
(Abgerufen am 05.12.2021)
- [36] AWS Marketplace
<https://aws.amazon.com/mp/marketplace-service/overview/> (Abgerufen am 06.12.2021)
- [37] Erin Carlson and Alea Whitman. Getting Started: Tracking AWS Cost Management Metrics
<https://aws.amazon.com/blogs/aws-cloud-financial-management/getting-started-tracking-aws-cost-management-metrics/>
(Abgerufen am 06.12.2021)
- [38] AWS Support - Benutzerhandbuch
https://docs.aws.amazon.com/de_de/awssupport/latest/user/support-ug.pdf#trusted-advisor
(Abgerufen am 07.12.2021)

-
- [39] AWS Support Plan Pricing
https://aws.amazon.com/premiumsupport/pricing/?nc1=h_ls
(Abgerufen am 09.12.2021)
- [40] Amazon Elastic Container Service Entwicklerhandbuch - Load Balancer-Typen
https://docs.aws.amazon.com/de_de/AmazonECS/latest/developerguide/ecs-dg.pdf#load-balancer-types
(Abgerufen am 09.12.2021)
- [41] 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 10.06.2020)
- [42] Definition von Ausgabe im Rechnungswesen
<https://wirtschaftslexikon.gabler.de/definition/ausgaben-31469#head1> (Abgerufen am 11.12.2021)
- [43] Bürgerlichen Gesetzbuch (BGB) § 630f
https://www.gesetze-im-internet.de/bgb/__630f.html
(Abgerufen am 08.12.2021)
- [44] SevDesk: Definition von Budgetplanung
<https://sevdesk.de/lexikon/budgetplanung/#budgetplanung-definition>
(Abgerufen am 28.11.2021)
- [45] Indeed:Cost Control Methods: Definitions and Examples
<https://www.indeed.com/career-advice/career-development/cost-control-methods>
(Abgerufen am 29.11.2021)
- [46] Ubuntu, delivered by Canonical:A business guide to hybrid/multi-cloud
https://ubuntu.com/engage/multi-cloud-business-guide?utm_source=google_ad&utm_medium=cpc&utm_campaign=7014K000000mSwp&gclid=Cj0KCQiAtJeNBhCVARIsANJUJ2Fb2Xp3WST3woFmmI11ZfqsMTRzvLVld-B1PE0yKVxdhm4tgxMklwcB
(Abgerufen am 29.11.2021)

-
- [47] The NIST Definition of Cloud Computing
National Institute of Standards and Technology(NIST) <https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-145.pdf>
(Abgerufen am 09.12.2021)
- [48] IDC Business Value of AWS 2015
http://d0.awsstatic.com/analyst-reports/IDC_Business_Value_of_AWS_May_2015.pdf (Abgerufen am 22.10.2021)
- [49] Instagram: Wann verschwindet meine Instagram Story?
<https://help.instagram.com/1729008150678239> (Abgerufen am 08.12.2021)
- [50] Online Marketing: Definition von Instagram Story?
<https://onlinemarketing.de/lexikon/definition-instagram-story> (Abgerufen am 09.12.2021)
- [51] Hilfebereich Instagram: VII. Datenspeicherung, Deaktivierung und Löschung von Konten.
<https://help.instagram.com/519522125107875> (Abgerufen am 12.12.2021)
- [52] Raj Bala, Bob Gill, Dennis Smith, Kevin Ji, David Wright.
Magic Quadrant für Cloud-Infrastruktur und Plattform-Services
<https://www.gartner.com/technology/media-products/reprints/AWS/1-271W10SP-DEU.html>
(Abgerufen am 23.09.2021 / Veröffentlicht am 27. Juli 2021)
- [53] Definition von Customer Acquisition Cost (CAC)
<https://onlinemarketing.de/lexikon/definition-customer-acquisition-cost-cac>
(Abgerufen am 06.12.2021)
- [54] Definition von Freemium
<https://onlinemarketing.de/lexikon/definition-freemium> (Abgerufen am 06.12.2021)
- [55] Definition von Cost-per-Action (CPA)
<https://onlinemarketing.de/lexikon/definition-cost-per-action-cpa>
(Abgerufen am 06.12.2021)
- [56] LinkedIn: Listado de todos los Servicios de AWS
<https://www.linkedin.com/pulse/listado-de-todos-los-servicios-amazon-web-ser-C3%B1a-silva/?originalSubdomain=es> (Abgerufen am 18.11.2021)

-
- [57] LinkedIn Learning: AWS Controlling Cost by Lynn Langit
<https://www.linkedin.com/learning/aws-controlling-cost/aws-service-types?autoAdvance=true&autoSkip=false&autoplay=true&resume=false&u=79182202> (Abgerufen am 29.11.2021)
- [58] SAP: Definition von maschinellen Lernen
<https://www.sap.com/germany/insights/what-is-machine-learning.html>
(Abgerufen am 09.12.2021)
- [59] Medium: How TrueCar Saves 40% on AWS with EC2 Reserved Instances
<https://medium.com/driven-by-code/how-truecar-saves-40-on-aws-with-ec2-reserved-instances>
(Abgerufen am 02.12.2021)
- [60] Techterms Definition Metadata.
<https://techterms.com/definition/metadata>
(Abgerufen am 08.12.2021)
- [61] Plusserver: Kostenoptimierung in AWS
https://get.plusserver.com/hubfs/Assets/aws/a/Whitepaper-Kostenoptimierung-in-AWS-DE.pdf?utm_campaign=IoT&utm_medium=email&_hsmi=188763947&_hsenc=p2ANqtz--pG4zb_6horYqX3d0QDpUAzNYdJL51HEBdAtK3IQRBKUfR226JxBly6n2ILDtAmkmPwlib5J7qYjL10c6Fsl&utm_content=188763947&utm_source=hs_automation (Abgerufen am 29.11.2021)
- [62] TÜV Rheinland: Kurse zur Ausbildung von Cloud Architekten
<https://akademie.tuv.com/weiterbildungen/architecting-on-aws-489176?>
(Abgerufen am 29.11.2021)
- [63] Definition Horizontal Scaling
<https://www.techopedia.com/definition/7594/horizontal-scaling?ref=wellarchitected> (Abgerufen am 09.12.2021)
- [64] Definition Slack
<https://slack.com/intl/de-de/help/articles/115004071768-Was-ist-Slack-> (Abgerufen am 11.12.2021)
- [65] Stern, Adam, The Truth About Cloud Pricing
<https://www.forbes.com/sites/forbestechcouncil/2018/11/16/the-truth-about-cloud-pricing/>

-
- [the-truth-about-cloud-pricing/?sh=1f37bba42f33](#)
(Veröffentlicht am 16.11.2018)
- [66] Spot by NetApp, What are AWS spot instances?
<https://spot.io/what-are-ec2-spot-instances/>
(Abgerufen am 01.12.2021)
- [67] 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)
- [68] Statista: 2020 überholt die Cloud lokale Speichermedien
<https://de.statista.com/infografik/18231/cloud-vs-lokal-er-speicher/>
(Abgerufen am 18.11.2021)
- [69] 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)
- [70] 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)
- [71] 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)
- [72] 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)

-
- [73] Statista: Amazon ist die Nummer 1 in der Cloud
<https://de.statista.com/infografik/20802/weltweiter-marktanteil-von-cloud-in>
(Abgerufen am 08.12.2021)
- [74] Ashish G. Revar, Madhuri D. Bhavsar. Securing User Authentication using Single SignOn in Cloud Computing.
<https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6153227>
(Abgerufen am 05.12.2021)
- [75] YAML Org. Definition of YAML <https://yaml.org/> (Abgerufen am 12.12.2021)

Anhang

I Vorlage für einer Fakturierungsalarme in CloudWatch

JSON

```
    "SpendingAlarm": {
      "Type": "AWS::CloudWatch::Alarm",
      "Properties": {
        "AlarmDescription": { "Fn::Join": ["", [
          "Alarm if AWS spending is over $",
          { "Ref": "AlarmThreshold" }
        ]]],
        "Namespace": "AWS/Billing",
        "MetricName": "EstimatedCharges",
        "Dimensions": [{
          "Name": "Currency",
          "Value" : "USD"
        }],
        "Statistic": "Maximum",
        "Period": "21600",
        "EvaluationPeriods": "1",
        "Threshold": { "Ref": "AlarmThreshold" },
        "ComparisonOperator": "GreaterThanOrEqualToThreshold",
        "AlarmActions": [{
          "Ref": "BillingAlarmNotification"
        }],
        "InsufficientDataActions": [{
          "Ref": "BillingAlarmNotification"
        }]
      }
    }
  }
```

YAML

```
SpendingAlarm:
  Type: AWS::CloudWatch::Alarm
  Properties:
```



```
AlarmDescription:
  'Fn::Join':
  - ''
  - - Alarm if AWS spending is over $
    - Ref: AlarmThreshold
  Namespace: AWS/Billing
  MetricName: EstimatedCharges
  Dimensions:
  - Name: Currency
  Value: USD
  Statistic: Maximum
  Period: '21600'
  EvaluationPeriods: '1'
  Threshold:
  Ref: "AlarmThreshold"
  ComparisonOperator: GreaterThanThreshold
  AlarmActions:
  - Ref: "BillingAlarmNotification"
  InsufficientDataActions:
  - Ref: "BillingAlarmNotification"
```

¹⁰²AWS CloudFormation - Benutzerhandbuch. S.481.[34]

II Alarm für die monatliche Kosten anhand eines Budgets

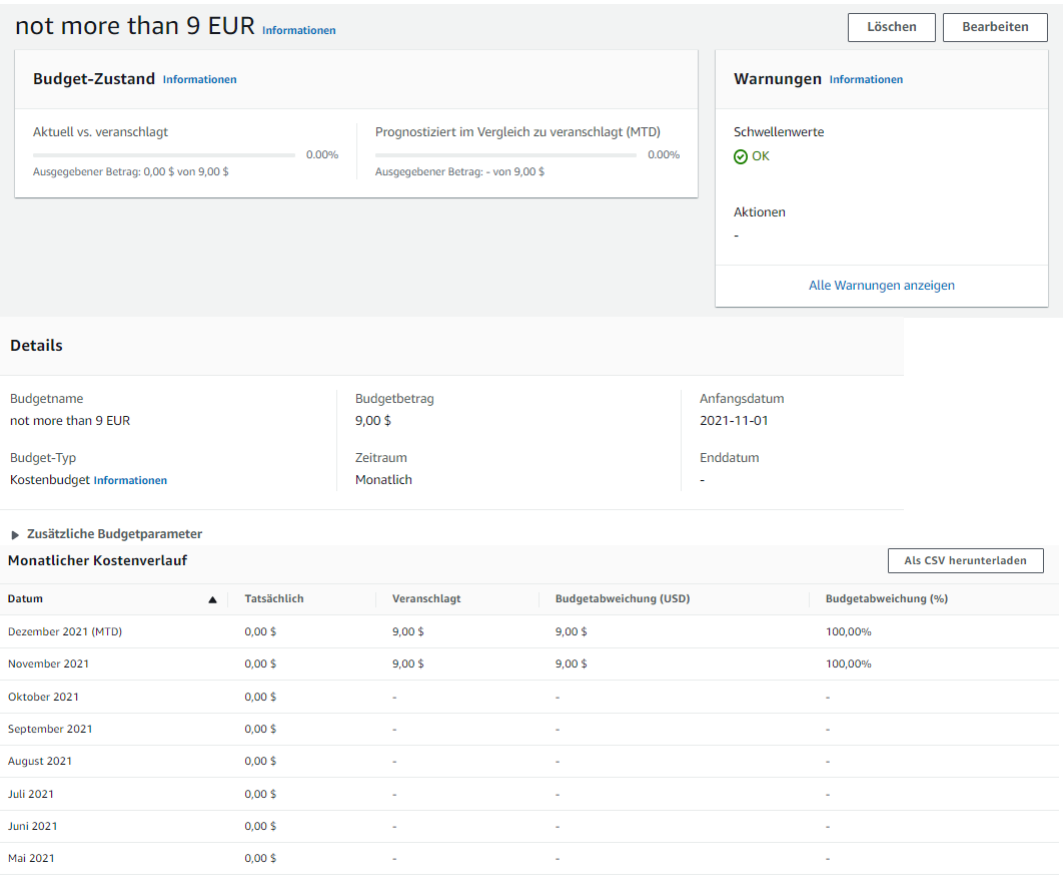


Abbildung 23
Eigene Darstellung von Test AWS-Konto.

[Rev Screenshot missing]

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 Zweifelsfalle 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.