



Fakultät Druck und Medien
Studiengang Medieninformatik

Bachelor Thesis
zur Erlangung des akademischen Grades Bachelor of Science

Evaluierung von Systemen zur Speicherung und Bereitstellung von Binärdaten im Kontext von Web Services

Gamze Isik

19. Juni 2023

Matrikelnummer: 39307

Bearbeitungszeitraum: 20. März - **19. Juni 2023**

Betreuer

Erstbetreuer: Prof. Martin Goik
Hochschule der Medien

Zweitbetreuer: Thomas Maier
Leomedia GmbH

Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbständig angefertigt habe. Es wurden nur die in der Arbeit ausdrücklich benannten Quellen und Hilfsmittel benutzt. Wörtlich oder sinngemäß übernommenes Gedankengut habe ich (mit Ausnahme dieser Erklärung) als solches kenntlich gemacht.

Ort, Datum

Unterschrift

Zusammenfassung

Das Ziel der vorliegenden Arbeit ist die Evaluierung eines geeigneten Systems zur Speicherung und Bereitstellung von Binärdaten im Kontext von Web Services. Dabei werden die Anforderungen wie Performance, Verfügbarkeit, Sicherheit und API Anbindung gestellt. Durch die Realisierung eines Prototyps anhand der ausgewählten Speicherlösung werden die Binärdaten durch sichere, zeitlich begrenzte URL's bereitgestellt. Folgende Fragen werden gestellt: Welches Speichersystem ist im Hinblick auf Kosten, Performance und Verfügbarkeit für die Persistenz von Binärdaten besonders geeignet? Wie kann man Daten durch sichere, zeitlich begrenzte URL's bereitstellen? Verschiedene Speichersysteme werden verglichen und anhand von Kostenkalkulationen bewertet. Durch die Auswertung des Vergleichs wird der Prototyp implementiert und Messungen auf Testdaten durchgeführt, welches die wissenschaftliche Arbeit stützt.*Das Ergebnis kann dazu genutzt werden, auf neue Speicherlösungen mit höherer Performance und Sicherheit mit akzeptablen Kosten umzusteigen.

Abstract

The aim of this thesis is to evaluate a suitable system for storing and providing binary data in the context of web services. Requirements such as performance, availability, security, and API integration are set. By implementing a prototype based on the selected storage solution, binary data is provided through secure, time-limited URLs. The following questions are addressed: Which storage system is particularly suitable for persisting binary data in terms of cost, performance, and availability? How can data be provided through secure, time-limited URLs? Different storage systems are compared and evaluated based on cost calculations. By evaluating the comparison, the prototype is implemented and measurements are taken on test data, which supports the scientific work. The result can be used to switch to new storage solutions with higher performance and security at acceptable costs.

Inhaltsverzeichnis

Akronyme	4
Abbildungsverzeichnis	5
Tabellenverzeichnis	5
1 Einleitung	6
1.1 Problemstellung und Motivation	6
1.2 Zieldefinition und Vorgehensweise	7
2 Speichersysteme	8
2.1 Arten von Speichersystemen	9
2.1.1 File Storage	10
2.1.2 Block Storage	11
2.1.3 Object Storage	12
2.2 Aktuelle Speichertechnologien im Markt	13
2.2.1 Eigenschaften	14
2.2.1.1 Sichere Speicherung	15
2.2.1.2 Hochverfügbarkeit	20
2.2.1.3 Kosten	23
2.2.1.4 Performance	27
2.2.1.5 API Anbindung	30
2.2.2 Bereitstellung der Dateien	32
2.3 Auswahl des Speichersystems	33
2.3.1 Kostenanalyse	33
2.3.2 Performance Analyse	33
2.3.3 Kalkulationsergebnisse	33
3 Prototypische Umsetzung	34
3.1 Überblick und Vorgehensweise	34
3.2 Technologien	34
3.3 Speicherung von Binärdaten	34
3.4 Bereitstellung der Binärdaten	34
3.5 Messung der Performance auf Testdaten	34
3.5.1 Messungsergebnisse	34
3.6 Zusammenfassung der Implementierung	34
4 Ergebnisse dieser Arbeit	35
4.1 Beschreibung und Funktionalität des Prototyps	35
4.2 Zusammenfassung der Ergebnisse	35
4.2.1 Kalkulationsergebnisse	35
4.2.2 Messungsergebnisse	35

5	Diskussion	36
5.1	Analyse und Interpretation der Ergebnisse	36
5.2	Bewertung des Prototyps	36
6	Fazit	37
6.1	Zusammenfassung der Ergebnisse	37
6.2	Beantwortung der Forschungsfrage	37
6.3	Potenzielle Anwendung des Prototyps	37
7	Danksagung	38
8	Literaturverzeichnis	39
9	Anhang	40
9.1	Prototyp	40
9.1.1	Github Link	40
9.1.2	Dokumentation	40
9.1.3	Code Snippets	40

Akronyme

Abb.	Abbildung
Anm.	Anmerkung
AWS	Amazon Web Services
dt.	deutsch
engl.	englisch
GCP	Google Cloud Platform
HTML	Hypertext Markup Language
URL	Uniform Resource Locator
USA	United States of America, dt. Vereinigte Staaten von Amerika

Tabellenverzeichnis

1 Einleitung

Das folgende Kapitel dient der Einführung in die Problemstellung, Motivation sowie Ziele und Vorgehensweisen der vorliegenden Arbeit.

1.1 Problemstellung und Motivation

Die steigende Menge an Binärdaten im Kontext von Web Services, die in verschiedenen Anwendungen generiert werden, stellt eine große Herausforderung dar. Dabei ist es von großer Bedeutung, dass diese Daten sicher, zuverlässig und schnell gespeichert und abgerufen werden können. Vor diesem Hintergrund stellen sich Fragen nach der Auswahl eines geeigneten Speichersystems, das die Anforderungen wie Performance, Verfügbarkeit, Sicherheit und API-Anbindung erfüllt. Zudem müssen Mechanismen bereitgestellt werden, um den Zugriff auf die Daten zu beschränken durch sichere, zeitlich begrenzte URL's.

Diese Bachelorarbeit richtet sich auf das Problem einer Full-Service-Ticketing Software „leoticket“, die vom Unternehmen Leomedia GmbH entwickelt wurde. Leomedia GmbH ist ein Unternehmen, das Software für Medienunternehmen wie Zeitungsverlage, Radiosender, Veranstalter, Künstler und Kulturvereine entwickelt. ^{leomedia-web}Leomedia (Bachelor- und Master-Themen [o.D.], [9])

Leoticket ist eines der vielen Produkte von Leomedia, dass Services wie Online-Kartenvorverkäufe, Abendkassen, den Einlass bei der Veranstaltung, Statistiken und Abrechnungen und die Planung der Veranstaltung realisiert. ^{leomedia-web}Leomedia (ebd.)

Das Problem liegt bei der Speicherung und Bereitstellung der Daten. Da es sich um Replikationen der Daten handelt, ist die Belastung des Systems hoch. Hohe Daten werden herumgeschoben. Die Bandbreite ist bei der Übertragung begrenzt. Ein weiteres Problem ist die Bereitstellung der Daten über Email Anhänge. Anhänge dürfen eine bestimmte Speichergröße nicht überschreiten. Wenn Ticketkäufer zehn oder 100 digitale Tickets kaufen, dann müssen diese über Email Anhänge bereitgestellt werden.

1.2 Zieldefinition und Vorgehensweise

Ziel dieser Arbeit ist die Realisierung eines Prototypen anhand der ausgewählten Speicherlösung die Binärdaten durch sichere, zeitlich begrenzte URLs bereitstellt. Dabei werden folgenden Fragen gestellt:

- Welches Speichersystem ist im Hinblick auf Kosten, Performance und Verfügbarkeit für die Persistenz von Binärdaten besonders geeignet?
- Wie kann man Daten durch sichere, zeitlich begrenzte URL's bereitstellen?

Im Rahmen der vorliegenden Bachelorarbeit werden verschiedene Arten von Speichersystemen untersucht, um die Forschungsfragen zu beantworten. Dabei erfolgt eine Analyse der aktuell verfügbaren Speichertechnologien auf dem Markt hinsichtlich ihrer Eigenschaften wie Sicherheit, Verfügbarkeit, Performance und Kosten. Bei der Berücksichtigung der Integration des Speichersystems in Software-Produkte wird auch die API-Anbindung des Speichersystems betrachtet. Zur sicheren Bereitstellung von Dateien werden zudem geeignete Cloud-Provider miteinander verglichen. Kosten- und Performance-Kalkulationen werden durchgeführt, um eine geeignete Speicherlösung auszuwählen. Die Ergebnisse werden anschließend ausgewertet.

Im Zuge der prototypischen Umsetzung werden die ausgewählten Technologien implementiert und Testdateien zur Verfügung gestellt. Nach der Durchführung von Messungen zur Performance auf Testdaten erfolgt eine Zusammenfassung der Implementierung.

Abschließend werden die Ergebnisse nochmals dargestellt und interpretiert sowie Schwächen und Grenzen des Prototyps aufgezeigt. Zur Einhaltung des roten Fadens der Arbeit werden die gestellten Forschungsfragen beantwortet und potenzielle Anwendungen des Prototyps aufgelistet.

2 Speichersysteme

Speichersysteme sind eine entscheidende Komponente der IT Infrastruktur eines Unternehmens. In der heutigen Zeit kann man von Speichersystemen kaum absehen, da Big Data immer an Wichtigkeit gewinnt. Sie bieten eine Möglichkeit, große Mengen an Daten zu speichern und zu verwalten, um den Zugriff und die Nutzung zu erleichtern. Es gibt eine Vielzahl an Speichersystemen, die für verschiedene Zwecke konzipiert sind. Durch die große Auswahl in der IT und die stetig anwachsende Innovation stellt sich die Frage, welche Speichersysteme sich für bestimmte Zwecke (Use Cases) eignen. Die Wahl des richtigen Speichersystems hängt von den Anforderungen des Unternehmens ab, wie zum Beispiel der Art der zu speichernden Daten, dem benötigten Zugriff und die Skalierbarkeit. Eine gründliche Analyse der Anforderungen und Kosten ist entscheidend, um die beste Lösung zu finden, die den Bedürfnissen des Unternehmens entspricht.

Im nachfolgenden Kapitel werden die unterschiedlichen Typen von Speichersystemen vorgestellt, wobei der Fokus auf den drei Speicherarten File-, Object- und Block Storage liegt. Im Anschluss daran erfolgt ein Vergleich von zwei Cloud-Providern in Bezug auf sicherer Speicherung, Hochverfügbarkeit, Performance, Kosten, API-Anbindung sowie der Dateibereitstellung. Auf Basis dieser Kriterien wird eine Entscheidung darüber getroffen, welcher Provider den Bedürfnissen von "Leoticket" entspricht. Hierbei fließen die Kosten- und Performance-Analysen mit in die Entscheidung ein.

2.1 Arten von Speichersystemen

Im folgenden Abschnitt werden die verschiedenen Arten von Speichersystemen vorgestellt, die für die Speicherung von digitalen Daten verwendet werden. Hierbei werden die drei gängigsten Speicherarten File-, Object-, und Blob Storage behandelt.

Die heutige IT-Landschaft bietet eine Vielzahl von Speichersystemen, die je nach Bedarf und Anforderungen ausgewählt werden können. Neben den traditionellen Speichermedien wie Festplatten und Bandlaufwerken gibt es heute auch verschiedene Arten von Speichersystemen, die in der Cloud oder als lokale Lösungen bereitgestellt werden können. Dazu gehören unter anderem File Storage, Object Storage und Blob Storage.

Jeder dieser Speicherarten hat ihre spezifischen Vor- und Nachteile und ist für bestimmte Anwendungsfälle besser geeignet als andere.

2.1.1 File Storage

File Storage, auch dateiebenen- oder dateibasierter Storage genannt ^{redHat-storage} RedHat: File storage, block storage, or object storage? (2018), [11], ist eine Speicherlösung bei der Dateien auf einem Dateisystem gespeichert werden.

Dieses System wird auch als hierarchischer Storage bezeichnet und gilt als das älteste und am weitesten verbreitete Datenspeichersystem für Direct und Network-Attached Storage. Dateisysteme organisieren Daten in hierarchischen Ordnern und Unterverzeichnissen, ähnlich wie in einem Dateiordner auf einem Computer. Dateien werden in der Regel auf einem Server oder einer Festplatte gespeichert und können von mehreren Benutzern gleichzeitig gelesen und geschrieben werden. Hierbei werden die Informationen in einzelnen Verzeichnisse abgelegt und können über den entsprechenden Pfad aufgerufen werden. Um dies zu ermöglichen, werden begrenzte Mengen an Metadaten ^{redHat-storage} RedHat genutzt, die dem System den genauen Standort der Dateien mitteilen, vgl. ^{redHat-storage} RedHat.

In der folgenden Abbildung wird die hierarchische Struktur des Dateispeichers visualisiert.

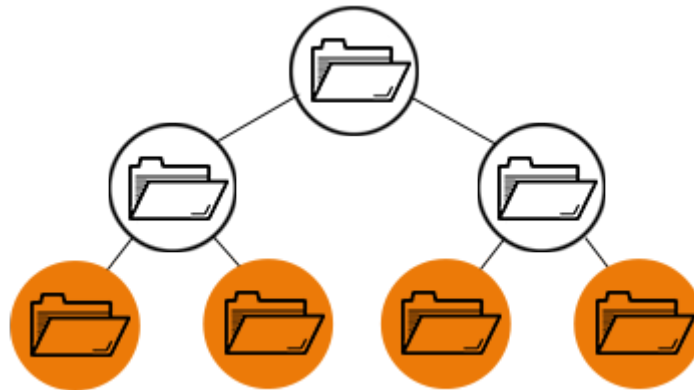


Abbildung 2.1.1: File Storage: Aufbau des Hierarchiesystems, ^{redHat-storage} RedHat

File Storage wird häufig in Unternehmen und Organisationen eingesetzt, um gemeinsame Dateiserver bereitzustellen oder Daten in Cloud-Speicherdiensten wie Dropbox oder Google Drive zu speichern. Auch wenn es von Betriebssystemen und Anwendungen gut unterstützt wird, kann die Performance und Skalierbarkeit von File Storage bei sehr großen Dateisystemen beeinträchtigt werden, was insbesondere bei stark frequentierten Anwendungen oder bei der Verarbeitung großer Datenmengen zum Problem werden kann.

Mit zunehmenden Datenvolumen erfordert das Skalieren von Dateispeichern das Hinzufügen neuer Hardwaregeräte oder den Austausch vorhandener Geräte durch solche mit höherer Kapazität. Dies kann im Laufe der Zeit teuer werden. „As data volumes expand, scaling file storage requires [...]“, ^{nx-fileScala} (Wahlman: Data Storage: Exploring File, Block, and Object Storage (2022), [13], Übersetzung des Autors)

Laut Wahlmann ^{nx-fileScala} (2022, Übersetzung des Autors) wird die Datenspeicherung bei zu vielen Daten nicht nur teuer, sondern auch unhandlich und zeitaufwändig. Der schnelle und einfache Zugriff auf jede einzelne Datei wird schwierig, wenn man zig Millionen von Dateien in Tausenden von Verzeichnissen auf Hunderten von Speichergrößen speichert.

2.1.2 Block Storage

Block Storage, auch genannt als Block-level Storage speichert Dateien auf SAN(Storage Area Networks) basierten Netzwerken oder auf Cloud-basierten Speicherumgebungen. Das System teilt Daten in Blöcke auf und speichert die separaten Teile jeweils mit einer eindeutigen Kennung, vgl. ^{ibm-topics}IBM: What is block storage? (o.D.), [6].

Daten werden in Blöcken auf dem Datenträger gespeichert, die unabhängig voneinander adressiert werden können. Jeder Block ist eine feste Größe, typischerweise im Bereich von einigen Kilobytes bis hin zu einigen Megabytes. Diese Blöcke können im System an jeder Stelle gespeichert werden.

Wenn auf Block Storage gespeicherte Daten abgerufen werden, verwendet das Server-Betriebssystem die eindeutige Adresse, um die Blöcke wieder zusammenzufügen und so die Datei zu erstellen. Der Vorteil besteht darin, dass das System nicht durch Verzeichnisse und Dateihierarchien navigieren muss, um auf die Datenblöcke zuzugreifen. Dadurch werden Effizienzen erzielt, da der Abruf von Daten schneller erfolgen kann, vgl. ^{ibm-storage}IBM: What is object storage? (o.D.), [7].

Typische Anwendungsbereiche des Block Storage sind Datenbanken, Virtualisierungsumgebungen und Anwendungen für Big Data-Analysen. Speicherung von strukturierten Daten wie Datenbanken, Virtuelle Maschinen und Betriebssysteme eignen sich besonders bei der Verwendung von Block Storage. Diese Art von Daten erfordert schnellen und direkten Zugriff auf bestimmte Bereiche des Speichers und muss häufig in Echtzeit ausgeführt werden. Block Storage eignet sich daher am besten für Anwendungen mit hohen Anforderungen an die Leistung und niedriger Latenzzeit.

2.1.3 Object Storage

Object Storage hat sich als Speichertechnologie in den letzten Jahren immer stärker etabliert und wird von vielen Unternehmen als Alternative zu traditionellen Speicherlösungen wie Block- oder File Storage angesehen. Die ersten Object Storage Systeme wurden bereits in den 1990er Jahren entwickelt, aber erst mit dem Aufkommen von Big Data, IoT und der Cloud-Nutzung hat es einen breiteren Einsatz gefunden. Heute bieten viele Cloud Provider wie Amazon Web Services (AWS) und Google Cloud Platform (GCP) Object Storage als einen ihrer Haupt-Cloud-Services an.

Object Storage ist für den Umgang mit großen Datenvolumen und unstrukturierten Daten entwickelt worden. Sie speichert Daten als eigenständige Objekte, die aus Daten und Metadaten bestehen und einen eindeutigen Identifier (UID) haben („Object storage (aka object-based storage) is a type of data storage used to [...]“, ^{dataCore-OS} Ivanov: File Storage vs. Object Storage: Understanding Differences, Applications and Benefits, What is Object Storage? (2020), [8], Übersetzung des Autors).

Im Gegensatz zu hierarchischen Systemen wie beim File Storage ist das Speichersystem flach strukturiert. Durch die einfache API Anbindung kann es mit vorhandenen Anwendungen integriert werden. Nutzer können detaillierte Informationen wie beispielsweise Erstellerangaben, Schlüsselwörter sowie Sicherheit- und Datenschutzrichtlinien hinterlegen. Diese Daten bezeichnet man als Metadaten.

Laut ^{nx-fileScala} Wahlman, 2022 ist Skalierbarkeit der Hauptvorteil, da bei der Speicherung von Petabyte und Exabyte alle Objekte in einem Namespace abgelegt werden. Selbst wenn dieser Namespace auf Hunderte von Hardwaregeräten und Standorten verteilt ist, können alle Objekte schnell abgerufen werden. Andere Vorteile von Objekt Storage beinhalten die Datenintegritätsprüfung, im Englischen bekannt als „erasure coding“ und die Datenanalyse.

Auch Object Storage hat seine Nachteile. Laut ^{redHat-storage} RedHat muss das Objekt nach der Speicherung bei Veränderung komplett neu überschrieben werden. Sie sind für traditionelle Datenbanken nicht geeignet, da das Schreiben von Objekten Zeit beansprucht und man sich mit der API auseinandersetzen muss, vgl. ^{redHat-storage} RedHat.

Insgesamt bietet Object Storage eine skalierbare und flexible Methode zur Speicherung von unstrukturierten Daten. Organisationen sollten jedoch die Vor- und Nachteile von Object Storage im Kontext ihrer spezifischen Anwendungsfälle abwägen, um eine fundierte Entscheidung über die beste Speichermethode zu treffen.

Da leoticket Daten wie Rechnungen und Tickets als Dateien speichern und abrufen abrufen, ist Object Storage die richtige Speicherart.

Für leoticket ist die Object Storage Variante die am Besten geeignetste Speicherlösung, da ...

2.2 Aktuelle Speichertechnologien im Markt

Die beiden größten Cloud-Speicheranbieter, Amazon Web Services (AWS) und Google Cloud Platform (GCP), bieten eine Vielzahl von Speicherlösungen an, die auf die Bedürfnisse von Unternehmen zugeschnitten sind.

In diesem Kapitel werden die aktuellen Speichertechnologien auf dem Markt untersucht, wobei der Fokus auf den Angeboten von AWS und GCP liegt. Um eine Vergleichsgrundlage zwischen AWS und GCP zu schaffen, werden die verschiedenen Aspekte wie sichere Speicherung, Hochverfügbarkeit, Leistung, Kosten, API Anbindung und die Bereitstellung der Dateien betrachtet. Dieses Vorgehen dient dazu, zu ermitteln, welche Speicherart sich am besten für welche Anforderungen eignet. Als Kontrast dazu wird das Open-Source-Objektspeichersystem MinIO betrachtet.

AWS bietet eine Reihe von Speicheroptionen an, darunter Amazon S3 (Simple Storage Service). Amazon S3 ist ein Object Storage-Service, der für die Speicherung und den schnellen Abruf von unstrukturierten Daten wie Videos, Fotos und Dokumenten ausgelegt ist.

Google Cloud Platform bietet ebenfalls eine Vielzahl von Speicherlösungen an, darunter Google Cloud Storage. Google Cloud Storage ist ein Object Storage-Service, der für die Speicherung von unstrukturierten Daten wie Bildern, Videos und Dokumenten ausgelegt ist.

Insgesamt bieten AWS und Google Cloud Platform eine Vielzahl von Speicherlösungen an, die auf die Bedürfnisse von Unternehmen zugeschnitten sind. Organisationen sollten jedoch die Vor- und Nachteile jeder Speicherlösung abwägen, um die beste Lösung für ihre spezifischen Anforderungen zu finden.

2.2.1 Eigenschaften

Im vorliegenden Abschnitt werden Amazon S3 und Google Cloud Storage in Bezug auf verschiedene Kriterien untersucht. Dabei werden zunächst die Eigenschaften von AWS S3 erläutert, gefolgt von einer Betrachtung von GC Storage. Ziel ist es, die Unterschiede zwischen den beiden Anbietern hervorzuheben und eine Entscheidungshilfe zu bieten, welche Plattform für die Anforderungen von „leoticket“ am besten geeignet ist. Die Auswahl der Kriterien erfolgt in Anlehnung an die spezifischen Anforderungen von „leoticket“.

2.2.1.1 Sichere Speicherung

Amazon S3

Viele Anbieter von Object Storage-Lösungen bieten integrierte Verschlüsselungsmöglichkeiten an, um sicherzustellen, dass Daten sowohl in Ruhe als auch in Bewegung geschützt sind. Dabei können unterschiedliche Verschlüsselungsmethoden zum Einsatz kommen. In Bezug auf die sichere Speicherung bieten sowohl AWS als auch GCP verschiedene Optionen für die Verschlüsselung von Daten. Die Sicherheit der gespeicherten Daten ist von entscheidender Bedeutung, um die Integrität und Vertraulichkeit der Daten zu gewährleisten. In diesem Unterabschnitt werden die verschiedenen Sicherheitsfunktionen von Amazon S3 untersucht.

IAM (Identity and Access Management) ist ein wichtiger Bestandteil von AWS und ermöglicht es Benutzern, Gruppen und Rollen zu erstellen, um den Zugriff auf S3 zu verwalten. Benutzer können individuelle Berechtigungen zugewiesen werden, während Gruppen und Rollen mehrere Benutzer mit denselben Berechtigungen zusammenfassen können. Auf S3-Buckets und Objekte kann man eine granuläre Zugriffssteuerung anwenden. Benutzer, Gruppen oder Rollen können so berechtigt werden, den Zugriff auf bestimmte Buckets und Objekte zu beschränken oder zu erlauben. "Beim Erteilen von Berechtigungen in Amazon S3 entscheiden Sie [...]" ^{aws-iam-s3} AWS: Security: Identity and Access Management (2023), [2]

Eine weitere wichtige Sicherheitsfunktion von Amazon S3 ist die Datenverschlüsselung. S3 bietet eine Vielzahl von Verschlüsselungsoptionen für die serverseitige und clientseitige Verschlüsselung. Da für leoticket die clientseitige Verschlüsselung nicht in Frage kommt, liegt der Fokus auf der serverseitigen Verschlüsselung. Es gibt drei Methoden, darunter die serverseitige Verschlüsselung mit Amazon S3-verwalteten Schlüsseln (SSE-S3), mit KMS-verwalteten Schlüsseln (SSE-KMS) und die SSE-C. Diese Optionen ermöglichen es Benutzern, die Verschlüsselung auf ihre spezifischen Anforderungen abzustimmen und so die Sicherheit der Daten zu gewährleisten.

Laut ^{aws-iam-s3} AWS nutzen Buckets standardmäßig die SSE-S3 Methode. Für die Verschlüsselung wird die 256-bit Advanced Encryption Standard (AES-256) verwendet. Seit dem 5. Januar 2023 sind alle neu erstellen Buckets auf SSE-S3 ausgelegt. Alle neuen Objekte sind beim Hochladen automatisch verschlüsselt, ohne weitere Zusatzkosten und keine Einbuße der Leistung.

Die Serverseitige Verschlüsselung schützt die Daten at rest. Amazon S3 verschlüsselt jedes Objekt mit einem eindeutigen Schlüssel. Als zusätzliche Sicherheitsmaßnahme werden diese eindeutigen Schlüssel mit einem weiteren Schlüssel verschlüsselt, welches in regelmäßigen Abständen rotiert wird, vgl. ^{aws-iam-s3} ebd.

Amazon S3 stellt auch die SSE-KMS als Auswahl zur Verfügung. AWS-KMS ist ein Dienst, das ein Schlüsselverwaltungssystem zur Verfügung stellt. Es verschlüsselt die Objekt Daten und speichert die S3 Checksum, das im Objekt Metadaten steckt, in verschlüsselter Form. Man kann die SSE-KMS in der AWS Management Konsole oder über die AWS KMS API verwalten. Jedoch gibt es zusätzliche Kosten bei der Verwendung der Methode. Dazu mehr im Kosten Abschnitt. Bei der Nutzung von AWS-SSE gibt es zwei Methoden. Einmal die AWS managed key oder die customer managed key. Es unterstützt die „envelope encryption“. Das bedeutet, dass die Schlüssel für die Daten durch einen Master Key verschlüsselt wird. Dies erleichtert die Verwaltung der Schlüssel.

Bei der AWS managed key Variante wird automatisch ein Schlüssel erstellt, sobald ein Objekt in ein Bucket hochgeladen wird. Dieser generierte Schlüssel wird dann für die Ver- und Entschlüsselung der Daten verwendet. Wenn man einen eigenen Schlüssel bei KMS verwenden möchte, dann erstellt man zuerst einen symmetrischen Key vor der KMS Konfiguration. Bei der Bucket Erstellung kann man anschließend den selbst-erstellten Key angeben. Die Nutzung von Customer Managed Keys hat einige Vorteile, die den Anforderungen von leoticket entspricht. Selbsterstellte Schlüssel bieten mehr Flexibilität und Kontrolle. Man kann sie selber erstellen, rotieren und deaktivieren. Hinzufügend kann man auch Zugriffskontrollen und Auditierung für den Schutz der Daten konfigurieren.

Wenn man die SSE-KMS Variante aussucht, kann man auch die S3 Bucket Key Funktion aktivieren. Dies kann die Request Kosten bis zu 99 Prozent reduzieren, indem die Request Traffic von Amazon S3 zu AWS KMS reduziert wird. Durch die Aktivierung der S3 Bucket Key für einen Bucket werden unique data keys für die Objekte im Bucket erstellt. Diese Bucket Keys werden für eine bestimmte Zeit verwendet, welches Abrufe zu Amazon S3 nach AWS KMS reduziert um Verschlüsselungsoperationen durchzuführen.

Zuletzt gibt es noch die SSE-C (server-side encryption with customer-provided keys). Bei der SSE-C stellt der Customer seinen eigenen Schlüssel zur Verfügung. Dieser Schlüssel wird nicht von Amazon S3 gespeichert, im Gegensatz bei AWS KMS. Amazon S3 übernimmt mit dem bereitgestellten Schlüssel die Datenverschlüsselung beim Schreiben sowie die Datenentschlüsselung beim Zugriff auf Objekte. Amazon S3 entfernt anschließend den Schlüssel aus dem Speicher. Da Amazon S3 den Schlüssel nicht speichert, speichert er den zufällig generierten HMAC (Hash-based Message Authentication Code) Wert vom Encryption Key um zukünftige Requests zu validieren.“Note: Amazon S3 does not store the encryption key that you provide. [...]”, ^{aws-sse-c} AWS: Using server-side encryption with customer-provided keys (SSE-C) (2023), [3].

Object Ownership ist eine weitere wichtige Sicherheitsfunktion von Amazon S3. Mit Object Ownership können Benutzer oder Gruppen die Eigentümerschaft von Objekten in S3-Buckets besitzen. Dies bedeutet, dass nur autorisierte Benutzer die Berechtigung haben, Objekte zu löschen oder zu ändern, was die Sicherheit der Daten erhöht. „S3 Object Ownership ist eine Einstellung auf Amazon-S3-Bucket-Ebene, mit der Sie Zugriffskontrolllisten (ACLs) deaktivieren und das Eigentum an jedem Objekt in Ihrem Bucket übernehmen können, [...]”. ^{aws-iam-s3} AWS: Security: Identity and Access Management (2023), [2]

AWS empfiehlt die ACL (Access Control List) auf Bucket-Ebenen deaktiviert zu lassen. Alle Objekte eines Buckets gehört so dem Bucket Owner. Laut [AWS](https://docs.aws.amazon.com/de-de/AmazonS3/latest/userguide//access-control-overview.html) verfügt Object Ownership über drei Einstellungen, mit denen man die Eigentümerschaft von Objekten steuern kann.

Einstellung	Gilt für	Auswirkung auf Object Ownership	Auswirkungen auf ACLs	Hochladen akzeptiert
Bucket-Eigentümer erzwungen (empfohlen)	Alle neuen und bestehenden Objekte	Bucket-Eigentümer besitzt jedes Objekt.	ACLs sind deaktiviert und wirken sich nicht mehr auf die Zugriffsberechtigungen für Ihren Bucket aus. Anfragen zum Festlegen oder Aktualisieren von ACLs schlagen fehl. Anfragen zum Lesen von ACLs werden jedoch unterstützt. Bucket-Eigentümer hat das volle Eigentum und die volle Kontrolle. Der Objekt-Writer hat nicht mehr das volle Eigentum und die volle Kontrolle.	Uploads mit ACLs mit vollem Zugriff des Bucket-Eigentümers oder Uploads, die keine ACL angeben
Bucket-Eigentümer bevorzugt	Neue Objekte	Wenn ein Objekt-Upload die bucket-owner-full-control vordefinierte ACL beinhaltet, gehört dem Bucket Eigentümer das Objekt. Objekte, die mit anderen ACLs hochgeladen wurden, gehören dem Schreibkonto.	ACLs können aktualisiert werden und können Berechtigungen erteilen. Wenn ein Objekt-Upload die bucket-owner-full-control vordefinierte ACL enthält, hat der Bucket-Eigentümer Vollzugriff und der Objekt-Writer hat keinen Vollzugriff mehr.	Alle Uploads
Objekt-Writer (Standard)	Neue Objekte	Der Objekt-Writer besitzt das Objekt.	ACLs können aktualisiert werden und können Berechtigungen erteilen. Der Objekt-Writer hat vollen Kontrollzugriff.	Alle Uploads

Abbildung 2.2.1: Einstellungen für Object Ownership, <https://docs.aws.amazon.com/de-de/AmazonS3/latest/userguide//access-control-overview.html>

Laut [AWS](https://docs.aws.amazon.com/de-de/AmazonS3/latest/userguide//access-control-overview.html) zeigt die obige Tabelle die Auswirkungen, die jede Einstellung für Object Ownership auf ACLs, Objekte, Objekteigentümer und Objekt-Uploads hat.

Logging ist ein weiterer Aspekt der Amazon S3-Sicherheit. S3 bietet verschiedene Logging-Optionen, darunter Bucket Logging und Object-Level Logging, die es Benutzern ermöglichen, Zugriffe auf S3-Objekte aufzuzeichnen und zu überwachen. Diese Funktionen sind entscheidend, um Compliance-Anforderungen zu erfüllen und verdächtige Aktivitäten zu erkennen. AWS bietet eine Vielzahl von tools zur Überwachung der Amazon S3 Ressourcen. Darunter die Amazon CloudWatch Alarms, AWS CloudTrail Logs, Amazon S3 Access Logs und die AWS Trusted Advisor.

Google Cloud Storage

Auch Google Cloud bietet eine Vielzahl an Sicherheitsfunktionen, um sicherzustellen, dass Daten in der Cloud sicher gespeichert und geschützt sind. Unter anderem die Datenverschlüsselung. Google Cloud bietet genau wie Amazon S3 Serverseitige Verschlüsselung an. Hier unterscheidet man zwischen Google-managed Keys, Customer-managed encryption keys und die Customer-supplied encryption keys. Google-managed encryption ist die Standard Verschlüsselungsoption von Google Cloud. Cloud Storage verschlüsselt die Daten serverseitig, bevor die Daten auf die Festplatte geschrieben werden.

Die Standard Variante verwaltet für den Nutzer die Encryption Keys in ihrem eigenen Key Management System. Auch Google verwendet für die Verschlüsselung die AES-256 wie AWS. Als Nutzer muss man bei dieser Variante keine Einstellungen berücksichtigen. Daten werden automatisch verschlüsselt beim Abruf, vgl. ^{gcp-encrypt}Storage: Google-managed encryption keys (2023), [12].

Wenn man mehr Kontrolle über die Schlüssel haben möchte, dann gibt es die Customer-managed encryption Option. Die Schlüssel werden von Cloud KMS (Cloud Key Management Service) erstellt und verwaltet. Diese Schlüssel werden vom Nutzer anschließend extern oder in einem HSM Cluster gespeichert. Customer Keys kann man für individuelle Objekte benutzen oder einen Bucket so einstellen, dass er einen Standard Key für alle neuen Objekte verwendet. Der erstellte Schlüssel wird für die Verschlüsselung der Objektdaten, die CRC32C Checksum des Objekts und für die MD5 Hash verwendet. Als zusätzlichen Schutz gibt es die Service Accounts, auch genannt als service accounts. Diesem Agent kann man bestimmte Rechte geben, sodass es Zugriff auf den gewünschten Encryption Key hat, um damit Objekte verschlüsseln zu können.

Zuletzt gibt es noch die Customer-supplied encryption keys. Als zusätzliche Sicherheitsebene zu Google-managed encryption keys können Nutzer ihren eigenen AES-256 Encryption Key bereitstellen, welches als Base64 encoded ist. Bei dieser Variante wird der Schlüssel nicht von Google Cloud gespeichert oder verwaltet. Die Nutzer müssen diese Schlüssel selber verwalten und speichern. Um zukünftige Requests zu validieren speichert Google Cloud einen kryptografischen Hash vom Schlüssel. Jedoch kann der Encryption Key nicht aus dem Hash regeneriert werden oder den Hash nicht zum Entschlüsseln anwenden.

Des Weiteren gibt es auch in Google Cloud Zugriffskontrolleinstellungen, mit der Nutzer genau steuern können, wer auf ihre Daten zugreifen kann. Den Zugriff kann man auf Bucket- und Objektebene steuern und Benutzern und Gruppen bestimmte Rollen zuweisen. Dabei wird zwischen Uniform und Fine-grained unterschieden. Uniform Access Control (UAC) ermöglicht es, den Zugriff auf einen gesamten Bucket in Google Cloud Storage auf der Ebene von Rollen zuzuweisen. Dazu können verschiedene vordefinierte Rollen wie "Leser", "Schreiber" oder "Besitzer" verwendet werden, die festlegen, welche Aktionen ein Benutzer auf dem Bucket ausführen darf. UAC ist eine einfachere Methode zur Zugriffskontrolle, da die Zugriffsrechte auf Bucket-Ebene vergeben werden. Das bedeutet, dass alle Objekte in dem Bucket automatisch dieselben Zugriffsrechte haben.

Fine-Grained Access Control (FGAC) hingegen ermöglicht es, die Zugriffskontrolle auf die Ebene von Objekten oder sogar auf Teile von Objekten herunterzubrechen. Das bedeutet, dass jeder

Benutzer oder jede Gruppe individuelle Zugriffsrechte auf bestimmte Objekte oder Teile von Objekten haben kann. Mit FGAC kann man sehr granulare Zugriffssteuerungen implementieren. Es ist eine mächtige Methode, aber auch komplexer und zeitaufwändiger zu implementieren als UAC. Mit Cloud IAM können auch rollenbasierte Zugriffssteuerungen für Bucket- und Objektebene ähnlich wie bei AWS umgesetzt werden.

Weitere Funktionen die Google Cloud anbietet ist Objekt Versioning, Audit Logging, Bucket Locks, VPC Service Controls, Data Loss Prevention und Identity-Aware Proxy.

2.2.1.2 Hochverfügbarkeit

Amazon S3

Amazon S3 ist ein skalierbarer und hochverfügbarer Objekt Storage, der eine Verfügbarkeit von 99.99 Prozent garantiert. „Designed to provide 99.999999999 percent durability and 99.99 percent availability of objects over a given year.“, ^{aws-availability} AWS: Data Protection in Amazon S3 (2023), [1]

Dies wird durch die Verwendung von Multi-Availability Zone Architekturen erreicht, die eine automatische Replikation von Daten in verschiedenen physischen Standorten ermöglichen. Die Multi-Availability Zone Architektur von Amazon S3 basiert auf der Aufteilung von Daten in mehrere geografisch getrennte Verfügbarkeitszonen (AZs). Jede AZ besteht aus mehreren physischen Rechenzentren, die sich in einem geografisch getrennten Gebiet befinden. Jede AZ ist vollständig unabhängig und bietet eine hohe Redundanz und Verfügbarkeit. Wenn ein Benutzer ein Objekt in Amazon S3 hochlädt, wird es automatisch in mehrere AZs repliziert, um sicherzustellen, dass das Objekt auch bei Ausfällen in einer AZ weiterhin verfügbar ist. Im Falle eines Ausfalls einer AZ wird Amazon S3 automatisch die Anfragen auf eine andere AZ umleiten, um eine ununterbrochene Verfügbarkeit des Objekts sicherzustellen. Durch die Cross-Region Replikation werden Daten automatisch in andere AWS-Regionen repliziert. Dadurch kann eine hohe Verfügbarkeit der Daten im Falle eines Ausfalls einer gesamten AWS-Region gewährleistet werden.

Amazon S3 bietet auch eine hohe Verfügbarkeit von Metadaten, die für den Zugriff auf Objekte verwendet werden. Metadaten werden automatisch in mehrere AZs repliziert, um sicherzustellen, dass die Metadaten auch im Falle eines Ausfalls einer AZ verfügbar bleiben. Zusätzlich zu Multi-Availability Zone Architekturen verwendet Amazon S3 auch Fehlerkorrektur- und Erkennungsmechanismen wie CRC-Prüfungen, um die Integrität von Daten sicherzustellen und sicherzustellen, dass die gespeicherten Daten stets korrekt sind.

Durch Aktivierung der Versionierung wird jeder Objektversion, die in einem Amazon S-Bucket gespeichert ist, eine eindeutige Versions-ID zugewiesen. Wenn eine Objektversion versehentlich gelöscht oder überschrieben wird, kann die vorherige Version wiederhergestellt werden.

Um eine hohe Verfügbarkeit bereitzustellen, stellt Amazon S3 außerdem noch die S3-Transfer Acceleration zur Verfügung. Durch die Verwendung von Amazon S3-Transfer Acceleration können Benutzer die Übertragung großer Datenmengen beschleunigen, indem ein optimierter Netzwerkpfad genutzt wird. Dadurch kann die Verfügbarkeit der Daten verbessert werden, indem Verbindungsprobleme minimiert werden.

Um die Leistung von Amazon S3 zu überwachen und auf mögliche Probleme reagieren zu können wird CloudWatch bereitgestellt. Dadurch können Ausfallzeiten minimiert werden und herausfinden, an welchen Momenten die Verfügbarkeit am geringsten ist.

Insgesamt bietet Amazon S3 eine hochverfügbare und zuverlässige Speicherlösung, die durch Multi-Availability Zone Architekturen und Fehlerkorrekturmechanismen eine Verfügbarkeit von 99,99 Prozent gewährleistet.

Google Cloud Storage

Laut ^{gcp-sla}Cloud: Cloud Storage Service Level Agreement (SLA), [4] wird die Verfügbarkeit von Google Cloud Storage als Service Level Agreement (SLA) angegeben, das die garantierte Verfügbarkeit für den Dienst definiert. Gemäß dem aktuellen SLA von Google Cloud Storage beträgt die garantierte Verfügbarkeit für Multi-Regionaler Speicher 99,95 Prozent und für Regionale Speicher 99,9 Prozent. Diese Zahlen geben an, dass Google Cloud Storage darauf ausgelegt ist, eine sehr hohe Verfügbarkeit zu gewährleisten. In der folgenden Tabelle werden diese Werte nochmals belegt.

Covered Service	Monthly Uptime Percentage
Standard storage class in a multi-region or dual-region location of Cloud Storage	>= 99.95%
Standard storage class in a regional location of Cloud Storage; Nearline, Coldline, or Archive storage class in a multi-region or dual-region location of Cloud Storage	>= 99.9%
Nearline, Coldline, or Archive storage class in a regional location of Cloud Storage; Durable Reduced Availability storage class in any location of Cloud Storage	>= 99.0%

Abbildung 2.2.2: Verfügbarkeit der Speicherklassen gemäß Google Cloud Storage SLA ^{gcp-sla}<https://cloud.google.com/storage/sla>

Es ist jedoch zu beachten, dass die tatsächliche Verfügbarkeit von Google Cloud Storage von mehreren Faktoren abhängt, einschließlich der spezifischen Konfiguration, dem Datenzugriffsmuster, der Netzwerkverfügbarkeit und anderen betrieblichen Variablen. Es ist daher möglich, dass die tatsächliche Verfügbarkeit in der Praxis leicht von der garantierten Verfügbarkeit abweicht.

Auch Google Cloud Storage bietet die Multi-Region Storage an. Dies ermöglicht die Speicherung von Daten in mehreren Regionen weltweit. Dadurch werden die Daten redundant repliziert und bleiben auch im Falle eines Ausfalls einer Region verfügbar.

Object Versioning bietet wie AWS die Möglichkeit Objektversionen beizubehalten. Dadurch können vorherige Versionen von Objekten wiederhergestellt werden, falls sie versehentlich gelöscht oder überschrieben werden.

Mit der Bucket- und Objekt Replikation können Daten automatisch in andere Regionen oder Projekte repliziert werden. Dies ermöglicht eine geografische Redundanz und verbessert die Verfügbarkeit der Daten.

Google Cloud Storage verwendet verteilte Speichersysteme eine redundante Infrastruktur, um eine hohe Datensicherheit- und Verfügbarkeit zu gewährleisten. Die Daten werden in mehreren Datenzentren repliziert, um Ausfälle zu vermeiden.

Ein weiterer Punkt ist die Monitoring und Fehlererkennung wie Stackdriver Monitoring, um die Leistung und Verfügbarkeit von google Cloud Storage zu überwachen und auf potenzielle Probleme

zu reagieren.

2.2.1.3 Kosten

In diesem Abschnitt werden die Kosten von Amazon S3 und Google Cloud Storage betrachtet. Dabei werden die allgemeinen Kosten erläutert für welche die Cloud Provider Gebühren fordern. Bei der Kostenanalyse werden anschließend die Daten von leoticket genommen und eine grobe Abschätzung der Kosten durchgeführt.

Amazon S3 setzt Gebühren für die Speicherung, Anforderungen und Datenabrufe, Datenübertragung, Verwaltung und Analyse, Replikation und S3 Objekt Lambda. Für die Speicherung ist die Gebühr abhängig von der Größe der Objekte, der Speicherdauer während des Monats und der Speicherklasse. Es gibt verschiedene Speicherklassen zur Auswahl, die für verschiedenen Use Cases geeignet sind. Das wären die S3 Standard, S3 Intelligent-Tiering, S3 Standard – Infrequent Access, S3 One Zone – Infrequent Access, S3 Glacier Instant Retrieval, S3 Glacier Flexible Retrieval (früher S3 Glacier) und S3 Glacier Deep Archive. S3 Glacier wird in dieser Arbeit nicht behandelt, da leoticket keine Anwendung für diese Speicherklasse hat. Es werden die Speicherklassen S3 Standard, S3 Intelligent-Tiering, S3 Standard - Infrequent Access und S3 One Zone - Infrequent Access untersucht.

Um das Ganze zu veranschaulichen, werden die Preise durch die untere Tabelle präsentiert.

	Standard	Intelligent Tiering	Standard-IA	One Zone - IA
Speicher/GB im Monat	0.0245 USD	0.0245 USD	0.0135 USD	0.0108 USD
PUT-, COPY-, POST-, LIST-Anforderungen (Pro Anfrage)	0.0000054 USD	0.0000054 USD	0,00001 USD	0,00001 USD
GET-, SELECT- und alle anderen Anforderungen (pro Anfrage)	0,00000043 USD	0.00000043 USD	0,000001 USD	0,000001 USD
Datenabrufe	Kostenlos	—	0.01 USD	0.01 USD
S3 Select zurückgegebene Daten	0.0008 USD	0.02 USD	0.01 USD	0.01 USD
Datenübertragung eingehend zu Amazon S3 aus dem Internet	Kostenlos			
Datenübertragungen aus Amazon S3 in das Internet	0.09 USD pro GB			

Abbildung 2.2.3: Übersicht der Kosten der AWS S3 Speicherklassen

In der obigen Tabelle werden die Standard Preise von den verschiedenen Speicherklassen von Amazon S3 dargestellt. Die One Zone -IA hat dabei die niedrigsten Kosten im Vergleich zur Standard Speicherkategorie. Jedoch sind die Kosten für die Anfragen beim One zone - IA genauso wie beim Standard - IA höher als beim Standard und Intelligent Tiering. Zusätzlich werden bei Datenabrufen in den Fällen von Standard-IA und One Zone-IA Gebühren von 0.01 USD pro GB erhoben. So sind die Speicherklassen Standard und Intelligent Tiering für häufig abgerufene Daten, bei denen eine sofortige Verfügbarkeit und eine niedrige Latenzzeit erforderlich ist geeignet. Standard bietet eine hohe Leistung und ist ideal für Anwendungen, bei denen schneller Zugriff auf die Daten erforderlich ist. Die Intelligent Tiering eignet sich für Daten mit wechselndem Zugriffsverhalten. Sie analysiert das Zugriffsmuster der Objekte und verschiebt sie automatisch zwischen zwei Zu-

griffstiers. Dadurch kann man Kosten sparen, da man für den tatsächlichen Zugriff auf die Daten bezahlt und nicht für eine dauerhafte Speicherung in der teureren Speicherklasse.

Standard-IA und One Zone-IA bieten kostengünstigere Optionen für die Speicherung der Daten. Die Standard-IA ist für Daten geeignet, auf die weniger häufig zugegriffen wird, aber bei Bedarf schnell verfügbar sein müssen. Im Vergleich zur Standardklasse sind die Kosten für die Speicherung niedriger, aber es fallen zusätzliche Gebühren für den Datenabruf an. Wenn Daten von zwei Speicherklassen hergeschoben werden, dann fallen Kosten für den Speicherklassenwechsel an. Diese hängen von der Datenmenge ab, die verschoben werden.

Die One Zone-IA ist für Daten geeignet, auf die selten abgerufen werden, jedoch mit einer Einschränkung. Anders als bei den anderen Speicherklassen wird One Zone-IA nur in einer einzelnen Availability Zone in einer bestimmten Region gespeichert. Dies bedeutet, dass die Daten in einer einzigen AZ verfügbar sind und ein potenzieller Datenverlust auftreten kann, wenn diese AZ nicht verfügbar ist.

Bei der Replikation von Daten zahlt man Gebühren für die Speicherung in den ausgewählten Ziel-S3-Speicherklassen, für die primäre Kopie, für Replikations-PUT Anforderungen und für anwendbare Gebühren für den Speicherabruf mit seltenem Zugriff. Laut [aws preise](https://aws.amazon.com/de/s3/pricing/) zahlt man auch für den regionenübergreifenden Datentransfer OUT von S3 zu jeder Zielregion. Die Preise für Speicher- und PUT-Anfragen für die replizierte Kopie basieren auf den ausgewählten AWS-Zielregionen, während die Preise für Datenübertragungen zwischen den Regionen auf der AWS-Quellregion basieren. Wenn man S3 Replication Time Control nutzt, zahlt man eine Datenübertragungsgebühr für die Replikationszeitsteuerung sowie die Gebühren für S3-Replikationsmetriken, die zum selben Tarif abgerechnet werden wie angepasste Amazon-CloudWatch-Metriken. Die Kosten für die S3 Replication Time Control-Datenübertragung beträgt pro GB 0.015 USD. <https://aws.amazon.com/de/s3/pricing/>

Auch bei der Objekt Versionierung fallen Kosten an, wenn man ältere Versionen von Objekten beibehalten möchte. Die Kosten beziehen sich auf den zusätzlichen Speicherplatz, der für die Aufbewahrung der Versionen verwendet wird. Für das Tagging von benutzerdefinierten Metadaten bei Objekten selbst fallen keine direkten Kosten an. Allerdings können Kosten für das Abrufen von Tags über die S3-API (z.B. mit der ListObjects- oder GetObject-Operationen) anfallen, da dies als Datenabruf betrachtet wird und entsprechende Gebühren gemäß den AWS-Preisen Datenzugriff erhoben werden. Es ist zu beachten, dass die genauen Preise für Objektversionierung und Tagging von Amazon S3 sich im Laufe der Zeit ändern können.

Die genauen Preise und Kostendetails für S3 Standard IA können sich im Laufe der Zeit ändern. Es wird empfohlen, die aktuellsten Informationen auf der offiziellen AWS-Preiseseite oder im AWS-Kostenrechner zu überprüfen, um eine genaue Kosteneinschätzung zu erhalten.

Google Cloud Storage

Google Cloud Storage setzt Preise für die Komponenten Datenspeicher, Datenverarbeitung und Netzwerknutzung. Beim Datenspeicher ist wie bei Amazon S3 die Menge der in den Buckets gespeicherten Daten bedeutend. Preise für Speicher hängen von der Speicherklasse der Daten und dem Standort der Buckets ab. Bei der Datenverarbeitung werden Gebühren für die von Cloud Storage durchgeführten Verarbeitung, einschließlich Vorgangsgebühren, anwendbarer Abrufgebühren und Replikation zwischen Regionen erhoben. Bei der Netzwerknutzung werden Gebühren für die Menge der aus den Buckets gelesenen oder zwischen diesen verschobenen Daten erhoben. Google Cloud Storage bietet vier Speicherklassen an von denen drei untersucht werden. Die Standard, Nearline und Coldline. Diese drei Speicherklassen haben unterschiedliche Leistungseigenschaften und Kosten. Die Preise variieren je nach gewählter Speicherklasse.

Um den Vergleich zwischen Amazon S3 und Google Cloud Storage zu veranschaulichen wird eine Tabelle in der unteren Abbildung gezeigt.(siehe Abb. 2.2.3)

	Standard	Nearline	Coldline
Speicher/GB im Monat	0.023 USD	0.013 USD	0.006 USD
Vorgänge Klasse A pro 1000 Vorgänge	0.005 USD	0.01 USD	0.01 USD
Vorgänge Klasse B pro 1000 Vorgänge	0.0004 USD	0.001 USD	0.005 USD
Abrufgebühren	kostenlos	0.01 USD pro GB	0.02 USD pro GB
Ausgehender Traffic von Google Cloud ins Internet	0.12 USD pro 0-1TB monatlich		

Abbildung 2.2.4: Vergleich der Kosten von Google Cloud Storage Speicherklassen

Ähnlich wie bei Amazon S3 sind die Speicherkosten der Nearline und Coldline geringer als die Standard Speicherkasse. Dafür werden Gebühren für die Datenabrufe bei Nearline und Coldline erhoben. Die verschiedenen Speicherklassen sind für unterschiedliche Anwendungsfälle ähnlich wie bei Amazon S3 konzipiert. Die Standard Storage-Klasse bietet hohe Performance, niedrige Latenzzeiten und hohe Verfügbarkeit. Sie eignet sich gut für häufig genutzte Daten, bei denen schneller Zugriff und geringe Latenz von Bedeutung sind. Beispiele dafür sind aktive Anwendungen, Datenbanken oder Inhalte mit hohem Durchsatz. Die Nearline Storage-Klasse ist für seltener genutzte Daten konzipiert, auf die jedoch mit niedriger Latenzzeit zugegriffen werden muss. Es bietet niedrigere Speicherkosten als die Standard Storage, jedoch mit einer etwas längeren Zugriffszeit. Es eignet sich für Backup-Daten, Archivierung und lange Speicherung. Letztere Speicherkasse, die Coldline Storage ist für Daten ausgelegt, auf die selten zugegriffen wird und bei denen eine längere Zugriffszeit akzeptabel ist. Es bietet die niedrigsten Speicherkosten an. Sie eignet sich gut für langfristige Archivierung, Compliance-Daten und Backup-Daten.

Zusätzliche Kosten, die nicht in der Tabelle aufgelistet sind, sind die Kosten für die Replikation der Daten. Auf Wunsch können Nutzer in Cloud Storage Daten innerhalb einer Region, in Dual-Regionen oder Multiregionen replizieren. Zusätzlich zu den Daten, die in den hochgeladenen Objekten enthalten sind, werden benutzerdefinierte Metadaten auf die monatliche Speichernutzung

angerechnet. Für die benutzerdefinierten Metadaten "NAME:VALUE" erfasst Google Cloud Storage beispielsweise jedes Zeichen in NAME und VALUE als Byte, das mit dem Objekt gespeichert wird.

Beim vorzeitigen Löschen eines Objektes aus den Speicherklassen Coldline und Nearline werden Gebühren erhoben, da eine Mindestspeicherdauer angesetzt ist. Das vorzeitige Löschen beim Nearline beträgt pro GB pro Tag ca. 0.00043 USD und beim Coldline 0.0002 USD. Auch für Tags fallen Gebühren pro Monat von 0.005 USD an, dass für Buckets angehängt werden.

Auch hier fallen Kosten für die Objektversionierung an. Die Kosten setzen sich aus zwei Hauptkomponenten zusammen. Einmal den Speicher und die Anfragen. Jede Version eines Objekts belegt Speicherplatz im Bucket. Die Kosten basieren auf der Größe der Objekte und der Anzahl der Versionen, die gespeichert sind. Das Hochladen, Herunterladen oder Löschen von Objektversionen führt zu Anfragen an den Storage-Dienst. Für diese Anfragen können Gebühren anfallen, die sich nach der Anzahl der Anfragen richten. Es ist wichtig zu beachten, dass die Kosten für die Objektversionierung zusätzlich zu den regulären Kosten für die Speicherung und den Datenverkehr in Google Cloud Storage anfallen. Daher sollten die potenziellen Kosten der Objektversionierung in den Kalkulation einbezogen werden.

2.2.1.4 Performance

Amazon S3

Amazon S3 bietet verschiedene Funktionen und Dienste, die die Performance und Skalierbarkeit der Datenzugriffe verbessern.

Die S3 Transfer Acceleration ermöglicht eine schnellere Übertragung großer Dateien. Es nutzt das Amazon CloudFront-Netzwerk, um den Datentransfer zu beschleunigen. Durch die Verwendung von optimierten Netzwerkpfeilen und Caching-Techniken kann S3 Transfer Acceleration die Latenz reduzieren und die Übertragungsgeschwindigkeit erhöhen.

S3 Select ermöglicht das effiziente Abrufen von spezifischen Daten aus Objekten in Amazon S3. Anstatt ein gesamtes Objekt herunterladen zu müssen, können mit S3 Select nur die benötigten Daten abgefragt werden. Dies reduziert den Datenverkehr und beschleunigt den Abrufvorgang erheblich, insbesondere bei großen Daten.

S3 Cross-Region Replication ermöglicht die Replikation von Daten zwischen verschiedenen AWS-Regionen. Durch die Replikation in geografisch entfernte Regionen kann die Leistung verbessert werden, indem Daten näher an den Benutzer oder Anwendungen bereitgestellt werden. Dadurch werden niedrigere Latenzzeiten und eine bessere Verfügbarkeit erzielt.

Google Cloud Storage

Google Cloud Storage ermöglicht die Auswahl des geeigneten Speicherorts für Daten, um die Latenzzeiten zu minimieren. Man kann zwischen multi-regionalen Speicherstandorten oder regionalen Speicherstandorten wählen, um Daten näher an den Benutzer oder Anwendungen zu speichern und den Zugriff zu beschleunigen.

Durch die Integration mit Content Delivery Networks (CDNs) kann die globale Verteilung von Daten optimiert und die Bereitstellung beschleunigen. Das CDN stellt eine Zwischenspeicherung von Inhalten in Edge-Servern weltweit bereit, um den Zugriff auf Daten schneller und effizienter zu machen.

Bei der Performance spielt die Request Rate auch eine bedeutsame Rolle. Das bietet das Auto-Scaling von GCP an. Cloud Storage ist ein multi-tenant Service, was bedeutet, dass Benutzer die zugrunde liegenden Ressourcen gemeinsam nutzen. Um die gemeinsam genutzten Ressourcen optimal zu nutzen, haben Buckets eine anfängliche I/O Kapazität.”Cloud Storage is a multi-tenant service, meaning that users share the same set of underlying resources.[...]”, ^{[gcp-autoscale](#)} Cloud: Request rate and access distribution guidelines: Auto Scaling, [5] (Übersetzung aus Google Cloud)

Diese Kapazitäten betragen etwa 1000 Schreibzugriffsanfragen pro Sekunde für Objekte, einschließlich Hochladen, Aktualisieren und Löschen von Objekten. Es ist zu beachten, dass Cloud Storage auch eine kleinere Begrenzung für wiederholte Schreibvorgänge mit demselben Objektnamen hat. Auf Lesezugriffsanfragen pro Sekunde betragen sie bei 5000, einschließlich Auflisten von Objekten, Lesen von Objektdaten und Lesen von Objektmeldungen, vgl. ^{[gcp-autoscale](#)} [ebd.](#)

Wenn die Anfragehäufigkeit für einen bestimmten Bucket steigt, skaliert Cloud Storage automatisch und erhöht die I/O Kapazität für diesen Bucket, indem die Anfragelast auf mehrere Server verteilt wird.

Cloud Storage ermöglicht den parallelen Upload und Download von Daten, um die Übertragungsgeschwindigkeit zu maximieren. Es können mehrere Threads oder Prozesse verwendet werden, um Daten gleichzeitig hochzuladen oder herunterzuladen und so die Leistung zu verbessern.

Der Google Cloud Storage Transfer Service ermöglicht den schnellen und effizienten Transfer von großen Datenmengen. Daten können aus anderen Cloud-Speicherlösungen, On-Premise-Speichern oder öffentlichen Datensätzen in Google Cloud Storage übertragen werden, um Zeit und Bandbreite zu sparen.

Während die out-of-the-box Performance bereits stabil ist, gibt es einige Einstellungen und Empfehlungen von Google Cloud, um die Performance auf den Anwendungsfall zu optimieren. Um die Leistung messen zu können, bietet Google Cloud die "perfdiag"-Tool an. Der Autor McAnlis empfiehlt in seinem Block [gcp-blog](#) Optimizing your Cloud Storage performance: Google Cloud Performance Atlas dieses Tool zu verwenden, dass eine Reihe von Tests durchläuft, die die aktuelle Performance von einem Cloud Bucket protokolliert. Des Weiteren empfiehlt er die Nutzung des "gsutil"-Tools von Google Cloud, um kleinere Dateien schneller hochzuladen. Wenn man 20 000 Dateien, die jeweils 1kb groß sind, hochladen möchte, dann dauert der Overhead bei jedem individuellen Upload länger als die gesamte Uploadzeit gemeinsam. Deshalb sollte man Batch Operationen verwenden, da diese den Overhead reduzieren und die Leistung verbessern.

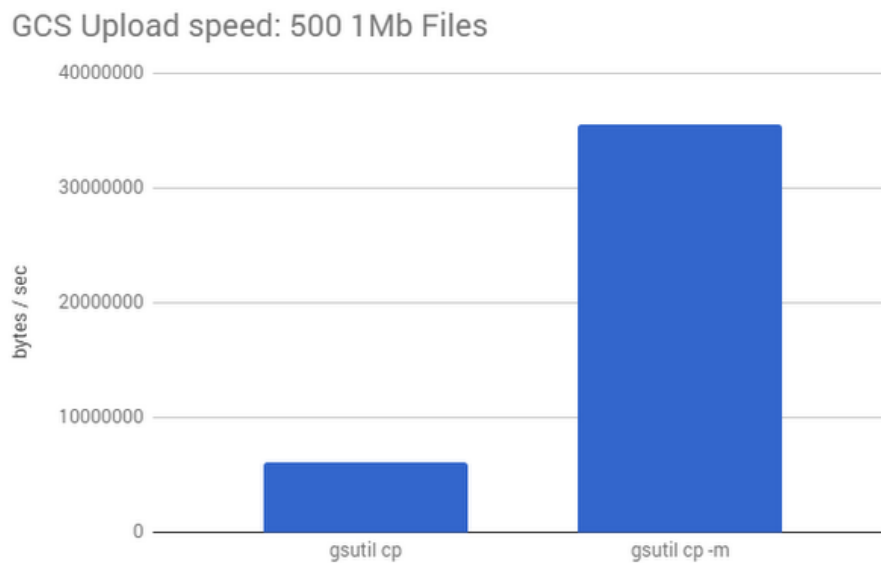


Abbildung 2.2.5: GCS Upload Geschwindigkeit, [gcp-blog](https://cloud.google.com/blog/products/) <https://cloud.google.com/blog/products/>

Es ist zu beachten, dass die Performance auch von anderen Faktoren abhängt, wie die Anwendungsarchitektur, die Netzwerkverbindung und die Implementierung in der Anwendung. Es können spezifische Konfigurationsoptionen genutzt werden, um die Leistung weiter zu optimieren, wie Caching, asynchrone Operationen oder die Nutzung von optimierten Bibliotheken oder Frameworks. Das gsutil Tool bietet die Option an Batch Operationen durchzuführen. Auf dem folgenden Diagramm wird ein Test mit 100 mal 200 000 Dateien mit individuellem Upload und Batch Upload angezeigt, das mit "gsutil -m cp" in ein Storage Bucket hochgeladen wird.

2.2.1.5 API Anbindung

Amazon S3

Die API-Anbindung von Amazon S3 ist umfangreich und bietet Entwicklern eine Vielzahl von Möglichkeiten zur Interaktion mit dem Speicherdienst. Amazon S3 bietet eine RESTful API (Application Programming Interface) sowie verschiedene SDKs (Software Development Kits) und Tools, die die Integration und Nutzung erleichtern.

Amazon S3 bietet eine RESTful API, die auf dem HTTP-Protokoll basiert. Mit dieser API können Entwickler HTTP-Anfragen wie GET, PUT, POST und DELETE verwenden, um auf Buckets und Objekte in Amazon S3 zuzugreifen, sie zu erstellen, zu lesen, zu aktualisieren und zu löschen.

Amazon S3 stellt offizielle SDKs für verschiedene Programmiersprachen bereit, darunter Java, Python, .NET, Ruby, Node.js und mehr. Diese SDKs bieten eine abstrakte Schnittstelle, um die Entwicklung von Anwendungen zu erleichtern und die Interaktion mit Amazon S3 zu vereinfachen. Darüber hinaus gibt es auch Drittanbieter-Tools und Open-Source-Bibliotheken, die die Integration mit Amazon S3 unterstützen. Beispielsweise MinIO

Amazon S3 ermöglicht die Durchführung von Batch-Operationen, um mehrere Objekte in einem einzigen API-Aufruf zu verarbeiten. Dies erleichtert die effiziente Verarbeitung großer Datenmengen und reduziert die Anzahl der API-Anfragen.

Die API-Anbindung von Amazon S3 erfordert eine geeignete Authentifizierung und Autorisierung, um sicherzustellen, dass nur autorisierte Benutzer auf die Daten zugreifen können. Dies erfolgt in der Regel mithilfe von Zugriffsschlüsseln, die über AWS Identity and Access Management (IAM) verwaltet werden.

Google Cloud Storage

Google Cloud Storage bietet sowohl JSON- als auch XML-APIs für die Interaktion mit dem Speicherdienst. Die JSON-API ist die bevorzugte Option und bietet eine moderne, leichtgewichtige Syntax für die Kommunikation mit dem Dienst. Neben der direkten API-Anbindung stellt Google auch das Google Cloud SDK zur Verfügung, das verschiedene Befehlszeilentools enthält, mit denen Sie auf Ihre Speicherressourcen zugreifen und sie verwalten können. Das SDK bietet auch eine Reihe von Clientbibliotheken für verschiedene Programmiersprachen.

Die API-Anbindung von Google Cloud Storage erfordert ebenfalls eine Authentifizierung und Autorisierung, um sicherzustellen, dass nur autorisierte Benutzer auf die Daten zugreifen können. Dies wird mithilfe von Google Cloud IAM (Identity and Access Management) verwaltet, das Zugriffsrichtlinien und Rollenverwaltung bietet.

2.2.2 Bereitstellung der Dateien

2.3 Auswahl des Speichersystems

In diesem Abschnitt wird durch eine Kosten-, und Performance Analyse die Speichersysteme ausgewertet und eine Auswahl getroffen. Die Kalkulationsergebnisse werden zum Schluss dargestellt.

2.3.1 Kostenanalyse

2.3.2 Performance Analyse

2.3.3 Kalkulationsergebnisse

3 Prototypische Umsetzung

3.1 Überblick und Vorgehensweise

3.2 Technologien

3.3 Speicherung von Binärdaten

3.4 Bereitstellung der Binärdaten

3.5 Messung der Performance auf Testdaten

3.5.1 Messungsergebnisse

3.6 Zusammenfassung der Implementierung

4 Ergebnisse dieser Arbeit

4.1 Beschreibung und Funktionalität des Prototyps

4.2 Zusammenfassung der Ergebnisse

4.2.1 Kalkulationsergebnisse

4.2.2 Messungsergebnisse

5 Diskussion

5.1 Analyse und Interpretation der Ergebnisse

5.2 Bewertung des Prototyps

6 Fazit

6.1 Zusammenfassung der Ergebnisse

6.2 Beantwortung der Forschungsfrage

6.3 Potenzielle Anwendung des Prototyps

7 Danksagung

8 Literaturverzeichnis

Internetquellen

- | | |
|------------------|--|
| aws-availability | [1] AWS: Data Protection in Amazon S3. (2023). Abgerufen am 14.05.2023. URL: https://docs.aws.amazon.com/AmazonS3/latest/userguide/DataDurability.html . |
| aws-iam-s3 | [2] AWS: Security: Identity and Access Management. (2023). Abgerufen am 11.05.2023. URL: https://docs.aws.amazon.com/de_de/AmazonS3/latest/userguide/access-control-overview.html . |
| aws-sse-c | [3] AWS: Using server-side encryption with customer-provided keys (SSE-C). (2023). Abgerufen am 12.05.2023. URL: https://docs.aws.amazon.com/AmazonS3/latest/userguide/ServerSideEncryptionCustomerKeys.html . |
| gcp-sla | [4] Google Cloud: Cloud Storage Service Level Agreement (SLA). Abgerufen am 17.05.2023. URL: https://cloud.google.com/storage/sla . |
| gcp-autoscale | [5] Google Cloud: Request rate and access distribution guidelines: Auto Scaling. Abgerufen am 17.05.2023. URL: https://cloud.google.com/storage/docs/request-rate . |
| ibm-topics | [6] IBM: What is block storage? (o.D.) Abgerufen am 07.05.2023. URL: https://www.ibm.com/topics/block-storage . |
| ibm-storage | [7] IBM: What is object storage? (o.D.) Abgerufen am 07.05.2023. URL: https://www.ibm.com/topics/object-storage . |
| dataCore-OS | [8] Mike Ivanov: File Storage vs. Object Storage: Understanding Differences, Applications and Benefits, What is Object Storage? (2020). Abgerufen am 07.05.2023. URL: https://www.datacore.com/blog/file-object-storage-differences/ . |
| leomedia-web | [9] GmbH Leomedia: Bachelor- und Master-Themen. (o.D.) URL: https://www.leomedia.de/unternehmen/abschlussarbeiten/ . |
| gcp-blog | [10] Colt McAnlis: Optimizing your Cloud Storage performance: Google Cloud Performance Atlas. Abgerufen am 18.05.2023. URL: https://cloud.google.com/blog/products/ . |
| redHat-storage | [11] RedHat: File storage, block storage, or object storage? (2018). Abgerufen am 06.05.2023. URL: https://www.redhat.com/en/topics/data-storage/file-block-object-storage . |
| gcp-encrypt | [12] Google Cloud Storage: Google-managed encryption keys. (2023). Abgerufen am 13.05.2023. URL: https://cloud.google.com/storage/docs/encryption/default-keys . |
| nx-fileScala | [13] Lauren Wahlman: Data Storage: Exploring File, Block, and Object Storage. (2022). Abgerufen am 06.05.2023. URL: https://www.nutanix.com/blog/block-storage-vs-object-storage-vs-file-storage . |

9 Anhang

9.1 Prototyp

9.1.1 Github Link

9.1.2 Dokumentation

9.1.3 Code Snippets