

Data Fabric Ansatz mit AWS



Whitepaper: Data Fabric Ansatz mit AWS

SEITE	THEMA
3	Vorwort
4	Einführung
6	Praxistest:
8	1. Data Lake und andere Datenspeicher
9	2. Data Warehouse
10	3. Data Engineering
10	4. Datenanalyse
11	5. Lösungen in Echtzeit
11	6. Business Intelligence
12	7. Data Science
12	8. Integration und Workflow-Orchestrierung
13	9. Sicherheit und Compliance
13	10. Kosten
14	Fazit
15	Über celfer

Vorwort

Microsoft und SAP haben mit ihren Data-Fabric-Konzepten neue Maßstäbe gesetzt: Die Plattformen vereinen verschiedene Data Services und ermöglichen eine nahtlose Integration und Orchestrierung von Daten.

Aber lässt sich das Konzept auch in anderen Umgebungen umsetzen – zum Beispiel in der Amazon Cloud? Wir machen den Praxistest!

Einführung

Wozu eine Datenplattform?

Daten sind überall. Sie sind in unseren Alltag integriert, an den Orten, an denen wir uns aufhalten: in unseren Bankverbindungen, beim Betreten von Geschäften, in unserem Suchverlauf und natürlich in allen Likes, Kommentaren und Ansichten von Social-Media-Beiträgen. Daten spiegeln sich auch in allen Kundentransaktionen, Produktivitätskennzahlen und Projektergebnissen unserer Teams wider. Wir alle sind Teil eines ständig wachsenden Daten-Stroms, in dem jede Interaktion von Bedeutung ist und das Potenzial hat, wichtige Entscheidungen zu beeinflussen.

Daten sind demnach in Hülle und Fülle vorhanden. Für sich allein sind diese Daten jedoch nicht aufschlussreich. Sie sagen uns nichts, solange sie nicht im richtigen Kontext interpretiert werden. Es ist daher die Aufgabe von Datenexperten, das Potenzial der Daten zu bewerten und die Schlüsselpunkte zu erkennen. Welche Details der Daten sind wichtig und welche Fragen können sie beantworten? Dieser Prozess beginnt mit dem Data Engineering. Zunächst müssen die Rohdaten gesammelt, aufbereitet, angemessen umgewandelt und sicher gespeichert werden, damit sie später adäquat genutzt werden können.

Bevor die Daten verwendet werden können, ist es jedoch wichtig, einen Anwendungsfall zu definieren. Die Formulierung der relevanten Fragestellungen und die Suche nach den richtigen Antworten sind der entscheidende Unterschied bei einer wertvollen Analyse. Datenanalysten versuchen in diesem Kontext, die Stakeholder, Führungskräfte oder Produktmanager in ihrem Entscheidungsprozess zu unterstützen. Dies geschieht durch die Erstellung von Dashboards und Berichten, die nicht nur die Daten anzeigen, sondern auch klare Antworten liefern und so eine solide Grundlage für Entscheidungen bieten.

Über die anfängliche Analyse hinaus vertiefen Data Scientists die Daten mithilfe fortschrittlicher statistischer Methoden, maschinellen Lernens und prädiktiver Analysen, um Muster aufzudecken, zukünftige Trends vorherzusagen und ein tieferes Verständnis der Daten zu ermöglichen. Mit ihrer Arbeit helfen sie den Teams nicht nur, das aktuelle Szenario zu verstehen, sondern auch, künftige Entwicklungen vorauszusehen und darauf zu reagieren.

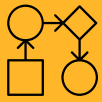
Einführung

Die Verwaltung des gesamten Datenlebenszyklus ist demnach eine komplexe Aufgabe, an der viele Datenexperten beteiligt sind. Die Zuverlässigkeit und Genauigkeit der Daten müssen unbedingt erhalten bleiben: Daten können es sich nicht leisten, falsch gehandhabt zu werden, beschädigt zu werden oder verloren zu gehen. Die Gewährleistung von Datenintegrität und -sicherheit ist von entscheidender Bedeutung. Gleichzeitig sind aber auch einfache und schnelle Zugriffsmöglichkeiten und Analysen gefragt.

Dieses Anforderungsprofil erfordert eine robuste Infrastruktur, die viele herkömmliche Systeme nicht bieten können. Große Software-Unternehmen haben diesen Bedarf erkannt und bieten daher nun Komplettlösungen an. So hat beispielsweise Microsoft mit Microsoft Fabric eine Plattform entwickelt, die zahlreiche Dienste in ein einzigartiges Produkt integriert und so die Komplexität der Datenverwaltung vereinfachen soll.

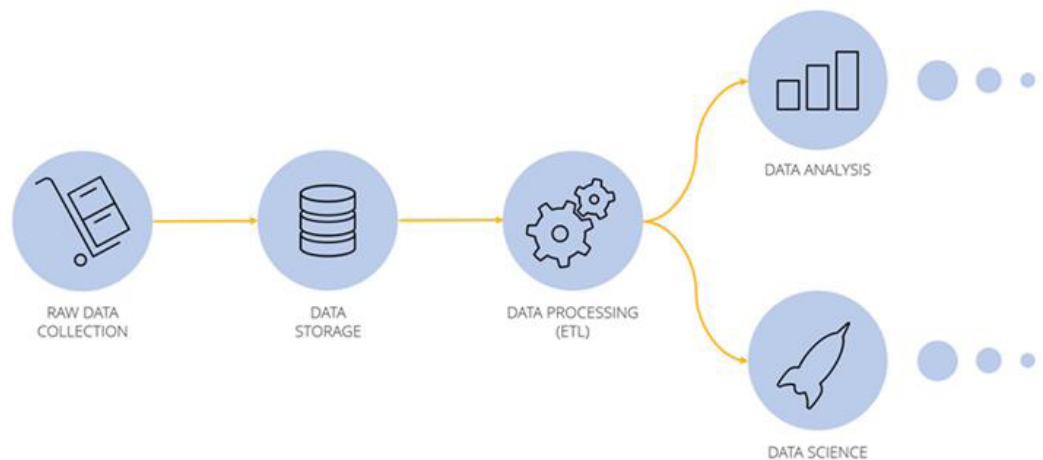
Aber was ist mit alternativen Lösungen wie AWS? Wir haben vor dieser Fragestellung untersucht, ob es mit den AWS-Services möglich ist, ein Produkt zu erstellen, das Microsoft Fabric entspricht.

Praxistest: Eignet sich AWS für den Fabric-Ansatz?

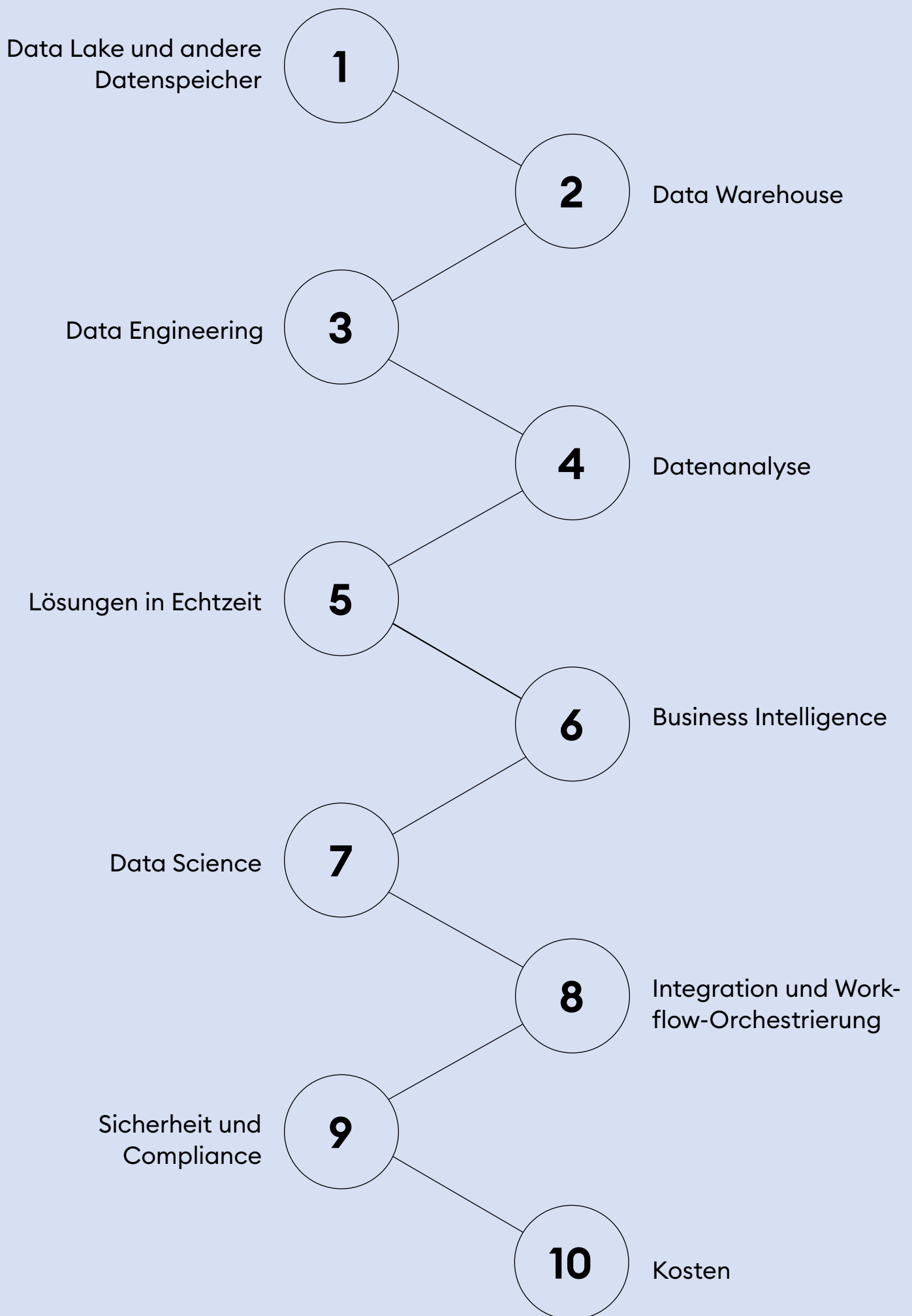


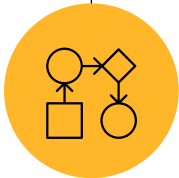
Wir haben eine komplette AWS-Datenpipeline-Infrastruktur in einem einzigen Produkt gekapselt, um das Management und die Orchestrierung in der Cloud zu testen. Diese Lösung wurde entwickelt, um Daten von ihrer rohen Form bis hin zu den anspruchsvollsten Stufen der Datenanalyse und prädiktiven Modellierung zu verarbeiten.

Da jeder Anwendungsfall seine eigenen Anforderungen hat, gewährleistet unsere einsatzfähige Lösung eine nahtlose Integration von der Datenaufnahme und -speicherung über komplexe ETL-Prozesse bis hin zur endgültigen Datenvisualisierung und den Modellen für maschinelles Lernen. Alles in einem Paket. Eine solche Flexibilität ist dank der Vielfalt der AWS-Services möglich, die die erforderliche Infrastruktur zur Unterstützung jedes Schritts eines Data Lifecycles bieten.



Allgemeine Data Pipeline





1. Data Lake und andere Datenspeicher

Die Grundlage einer robusten Datenarchitektur ist ein sicherer, skalierbarer und zugänglicher Speicher. AWS bietet Amazon Simple Storage Service (Amazon S3) als Data-Lake-Speicher an, der diese kritischen Anforderungen erfüllt. Die Robustheit und hohe Lebensdauer von Amazon S3 machen es zur idealen Umgebung für einen Data Lake und ermöglichen die Speicherung einer Vielzahl von Datentypen und -formaten aus unterschiedlichen Quellen. Der Dienst stellt sicher, dass die Daten sicher gespeichert werden und von mehreren AWS-Diensten aus leicht zugänglich und abrufbar sind.

Für eine bessere Einrichtung und Verwaltung von Data Lakes kann AWS Lake Formation verwendet werden, um die Sicherheit und Governance zu verbessern. AWS Lake Formation arbeitet mit Amazon S3 zusammen, um Daten zu katalogisieren und für die Analyse vorzubereiten und die Zugriffskontrolle und Compliance zu vereinfachen.

Zusätzlich zu einer Data-Lake-Lösung bietet AWS verwaltete Datenbankservices für Ihren speziellen Anwendungsfall. Amazon RDS erleichtert die Speicherung Ihrer relationalen Daten und bietet die Flexibilität, aus den gängigsten Datenbank-Engines wie MySQL, PostgreSQL oder Microsoft SQL Server zu wählen. Gleichzeitig bietet DynamoDB einen vollständig verwalteten NoSQL-Datenbankservice, der sich ideal für Ihre Anwendungen eignet, die hohe Leistung, Skalierbarkeit und Datenzugriff mit geringer Latenz erfordern.

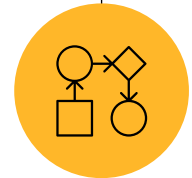
2. Data Warehouse

Für das Data Warehousing hat AWS einen speziellen AWS-Service: **Amazon Redshift** bietet eine vollständig verwaltete Data-Warehouse-Lösung im Petabyte-Bereich. Es wurde entwickelt, um Hochleistungsanalysen auf großen Datensätzen auszuführen, indem ein spaltenbasierter Datenspeicheransatz verwendet wird. Das verbessert die Leistung der meisten analytischen Abfragen.

Amazon Redshift bietet sowohl eine serverbasierte als auch eine **serverlose Option**, die es Ihnen ermöglicht, sich nicht um die Verwaltung der dahinterliegenden Infrastruktur kümmern zu müssen.

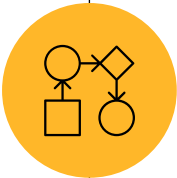
Darüber hinaus lässt sich Amazon Redshift mit Hilfe von **Amazon Redshift Spectrum** nativ in Ihren auf S3 aufgebauten Data Lake integrieren.

Parallel zu seinen Data-Warehousing-Kapazitäten hat AWS die Notwendigkeit erkannt, Daten nicht nur innerhalb von AWS, sondern auch aus externen Quellen zu migrieren. **AWS Database Migration Service** (DMS) ist ein AWS-Service, mit dem Sie Ihre Daten von mehreren lokalen Datenbanken, anderen Cloud-Plattformen oder zwischen AWS-Services migrieren können, wobei eine vielfältige Kombination von Datenquellen und Zielen unterstützt wird.



3

4



3. Data Engineering

Wie bereits erläutert, besteht das Data Engineering hauptsächlich aus dem Extrahieren von Daten aus Datenquellen, dem Durchführen der erforderlichen Transformationen und dem Laden der Daten in das benötigte Speichersystem, dem klassischen Batch-ETL (Extract, Transform, Load). **AWS Glue** ist ein ETL-Service, der diese Data-Engineering-Aufgaben erleichtert und eine Reihe von Konfigurationsoptionen bietet, die gleichzeitig serverlos sind. AWS Glue basiert auf **Apache Spark** und ermöglicht es Ihnen, Ihren Code in Python oder Scala zu schreiben und die Konfiguration und Ausführung Ihres Codes durch die Planung von Glue-Jobs zu verwalten.

AWS Glue lässt sich nahtlos in Amazon S3 sowie in andere Dienste wie RDS und Redshift integrieren und kann mit jedem JDBC-kompatiblen Datenspeicher verbunden werden. Die Glue Crawler extrahieren alle verfügbaren Metadaten und Schemata aus Ihren Datenquellen und stellen sie im **Glue Data Catalog** zur Verfügung. Dieser zentralisierte Katalog macht diese Datendetails für andere AWS-Services verfügbar und gewährleistet so eine konsistente und nahtlose Integration über die gesamte Plattform.

AWS Lambda wird als serverloser Rechenservice auch häufig zur Durchführung von Transformationsvorgängen für Daten verwendet, die in S3 gespeichert sind. Mit diesem Service können Sie Code als Reaktion auf Auslöser ausführen, ohne dass Sie einen Server im Hintergrund verwalten müssen.

4. Datenanalyse

Wenn es um die Durchführung einfacher und unkomplizierter Datenanalysen geht, ist **Amazon Athena** der richtige Dienst. Er ermöglicht Ihnen die einfache Analyse von Daten über Ad-hoc-SQL-Abfragen, ohne dass Sie eine Infrastruktur oder ein Data Warehouse verwalten müssen. Amazon Athena ist vielseitig, unterstützt mehrere Datenquellen und ermöglicht auch das Lesen von Daten aus Ihren S3-Buckets, was durch das Metadaten-Repository auf Basis des Glue Data Catalog erleichtert wird.

5. Lösungen in Echtzeit

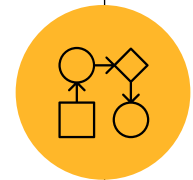
Wenn Ihre Anforderungen die Verarbeitung von Daten in Echtzeit voraussetzen, bietet AWS ebenfalls mehrere Dienste an. **Amazon Kinesis** bietet sich zum Beispiel in diesem Kontext mit einigen leistungstarker Tools an:

- Amazon Kinesis Data Streams kann Ihre Daten in Echtzeit aus mehreren verschiedenen Datenquellen in Ihrer AWS-Umgebung erfassen und speichern.
- Amazon Data Firehose ermöglicht es, Daten zu erfassen, umzuwandeln und in S3 oder Amazon Redshift zu laden und vereinfacht so Ihre Echtzeit-Analysen.
- Mit Amazon Kinesis Data Analytics können Sie Ihre Streaming-Daten mithilfe von SQL verarbeiten und analysieren. Dies erleichtert die Erstellung von Dashboards in Echtzeit oder die Erkennung von Anomalien in Ihrem Datenstrom.

Alternativ bietet AWS mit **AWS Managed Streaming for Kafka (MSK)** seine eigene native Apache Kafka-Implementierung als verteilte Streaming-Lösung mit hohem Durchsatz. MSK erleichtert die Verwaltung eines Kafka-Clusters und vereinfacht die Integration in Ihre bestehende AWS-Infrastruktur.

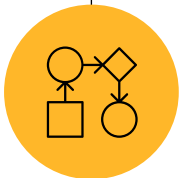
6. Business Intelligence

Für den Bereich Business Intelligence bietet AWS den Service **Amazon QuickSight** an. Es ermöglicht Ihnen, detaillierte und interaktive Berichte aus Ihren Daten zu erstellen, die unterschiedlichen Interessensgruppen Einblicke in Ihre Unternehmensdaten geben. Amazon QuickSight verfügt über integrierte Funktionen für maschinelles Lernen, mit denen Sie Ihre Berichte mit Prognosezahlen anreichern und Forecasts für eine bessere Entscheidungsfindung erstellen können. Amazon QuickSight ist sehr eng mit anderen AWS-Services verzahnt. So können Sie beispielsweise Ihre Berichte mühelos mit Daten erstellen, die direkt in Ihrem Data Lake gespeichert sind.



7

8



7. Data Science

Wenn Ihr Projekt spezielle Data-Science-Fähigkeiten erfordert, bietet AWS mehrere Dienste an, die für eine breite Palette von Aufgaben konzipiert sind. Zum Beispiel gibt es für die Modellierung von Forecasts den Service **Amazon Forecast**, für die visuelle Analyse **Amazon Rekognition**, für den Einsatz generativer KI und Large Language Models (LLMs) **Amazon Bedrock**, usw. Wenn Ihr spezieller Anwendungsfall noch von keinem der verfügbaren AWS-Services abgedeckt wird, bietet **AWS Amazon Sagemaker** an. Dieser Service ermöglicht es Ihnen, Ihre eigenen Modelle für maschinelles Lernen zu erstellen, zu trainieren und bereitzustellen.

8. Integration und Workflow-Orchestrierung

Wie Sie vielleicht schon bemerkt haben, sind die meisten AWS-Services eng mit Amazon S3 integriert, sodass Sie Ihre AWS-Infrastruktur auf Ihrem bereits vorhandenen Data Lake aufbauen können. S3 bietet Benachrichtigungen, mit denen Sie Ihre AWS-Workflows mit Hilfe von AWS-Services wie **Amazon EventBridge** oder **AWS Lambda** als Reaktion auf bestimmte Ereignisse auslösen können. Zum Beispiel, wenn neue Daten in Ihrem S3-Bucket ankommen.

Wenn Ihre AWS-Architektur eine einheitliche, aber dennoch entkoppelte Logik erfordert, können Services wie **Amazon Simple Queue Service (SQS)** oder **Amazon Simple Notification Service (SNS)** verwendet werden.

AWS Step Functions orchestriert und koordiniert Ihre Arbeitsabläufe und Prozesse innerhalb Ihrer AWS-Infrastruktur.

Die Integration des Aufbaus dieser gesamten AWS-Architektur kann durch die Verwendung unserer bereits erstellten **Terraform-Module** erleichtert werden. Diese Module wurden von uns entwickelt, um die Effizienz zu maximieren und den Aufbau Ihrer komplexen AWS-Infrastruktur zu beschleunigen, Redundanzen zu reduzieren und branchenübliche Best Practices zu fördern.

9. Sicherheit und Compliance

Das Thema Sicherheit nimmt AWS sehr ernst. Der Zugriff auf AWS-Ressourcen kann durch Richtlinien, **IAM-Benutzer** (Identity and Access Management) und IAM-Rollen genau kontrolliert werden, um sicherzustellen, dass nur autorisierte Personen die entsprechende Zugriffsstufe erhalten.

AWS Key Manage Service (KMS) bietet die Möglichkeit, Ihre AWS Encryption Keys zu erstellen und zu verwalten. KMS ist nahtlos in mehrere AWS-Dienste integriert und vereinfacht die Verschlüsselung Ihrer Daten.

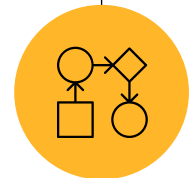
AWS Secrets Manager hilft Ihnen bei der Verwaltung und dem Schutz Ihrer vertraulichen Anmeldeinformationen, wie z.B. Datenbankzugängen oder API-Tokens. Durch die zusätzliche Aktivierung der Secret Rotation erhöhen Sie die Sicherheit Ihrer Informationen ohne manuelle Eingriffe.

Zur Identifizierung, Klassifizierung und zum Schutz sensibler Daten, die in Ihrem Data Lake in S3 gespeichert sind, verwendet **AWS Macie** Machine Learning um persönliche und vertrauliche Informationen zu erkennen und zu schützen.

10. Kosten

In AWS wird Ihnen Ihre Pipeline nicht als ein einziges Produkt in Rechnung gestellt, sondern jeder Service wird unabhängig abgerechnet. Dies geschieht in der Regel auf Grundlage seiner aktiven Nutzung. Dieser Preisansatz kann sehr vorteilhaft sein, da er es ermöglicht, die Ausgaben für jede Komponente in Ihrer AWS-Datenpipeline über den **AWS Cost Explorer** detailliert zu überwachen und zu verwalten. Diese Präzision erhöht die Transparenz und Anpassungsfähigkeit Ihrer Infrastruktur.

Wenn Sie jedoch sicherstellen müssen, dass Ihre Gesamtlösung innerhalb eines bestimmten Kostenrahmens bleibt, können Sie mithilfe von **AWS Budget** ein Limit festlegen.





AWS bietet eine breite Liste von Diensten, die die Entwicklung und Verwaltung einer vollständigen, einheitlichen Datenarchitektur ermöglichen, die absolut mit Microsoft Fabric vergleichbar ist. Das umfangreiche Angebot an AWS-Services bietet uns eine einzigartige Flexibilität, die richtigen Dienste auszuwählen und so zu kombinieren, dass sie Ihre spezifischen Bedürfnisse und Anforderungen erfüllen.

Dank der nahtlosen Integrationsmöglichkeit der AWS-Services und unserer Erfahrung bei der Entwicklung und Erstellung von Datenpipelines mit Infrastructure as Code (IaC) können wir Ihre spezifische Datenpipeline schnell entwickeln, anpassen und in einem einzigen, zusammenhängenden Paket für Sie konsolidieren. So können Sie sich direkt auf Ihre Geschäftsziele konzentrieren und müssen sich nicht mehr um die Verwaltung einer umfassenden Datenarchitektur kümmern.

Autorin:

Paula Guerra Toni ist Data Engineer bei der celver AG und beschäftigt sich neben den klassischen Aufgaben eines Data Engineers insbesondere mit Datenanalysen sowie der Architektur und Bereitstellung von Cloud-Umgebungen. In diesen Bereichen eignet sie sich kontinuierlich neues Wissen an und setzt dieses aktiv in ihrer Arbeit um. Paula verfügt über umfassende Projekterfahrung mit verschiedenen Tools und Technologien.

ÜBER UNS

Seit über 20 Jahren realisieren wir komplexe Planungs- und Analyselösungen auf Basis smarter Datenarchitekturen für namhafte Kunden unterschiedlichster Branchen. Die Grundlage bilden innovative Konzepte und Technologien, die sich von klassischen On-Premises-Lösungen hin zu agilen Cloud-Umgebungen unter Einbeziehung von Künstlicher Intelligenz entwickelt haben.

Das Ziel aber bleibt unverändert: Gemeinsam passgenaue Lösungen aufzubauen, die die Basis für datengestützte Entscheidungen bilden. Unser Team aus erfahrenen, zertifizierten Experten unterstützt dabei von der fachlichen Beratung über die Prozessdefinition, die komplette Realisierung und Integration in die Systemlandschaft bis hin zu Schulung, Roll-out und Change Management.

Als technologieunabhängiges Beratungsunternehmen sind wir Mitglied der UNITY Innovation Alliance.

KEYFACTS

- Über 20 Jahre Erfahrung
- >100 Mitarbeiter und >100 aktive Kunden
- Langjährige Kundenbeziehungen
- Technologieunabhängig
- Branchenexpertise: Handel & Konsumgüter, Fashion & Retail, Industrie & Logistik, Chemie & Pharma
- Mitglied der UNITY Innovation Alliance

SCHWERPUNKTE

- Smart Data Architecture
- DWH Modernization
- Data Management
- Data Platform
- Advanced Analytics
- Data Science
- Expert as a Service

KUNDEN

Aquila Capital Holding GmbH, BASF SE, Betty Barclay Group, Bionorica SE, GEPA Gesellschaft zur Förderung der Partnerschaft mit der Dritten Welt mbH, Hapag Lloyd AG, Lidl Dienstleistung GmbH & Co. KG, meinestadt.de GmbH, Olympus Europa SE & Co. KG, Optibelt GmbH, P & T Paper & Tea GmbH, RTG Retail Trade Group GmbH, Solvares Group GmbH, VTG AG, Walter Greif GmbH und Co. KG, uvm.

ANSPRECHPARTNER

Sie haben weitere Fragen zu unseren Services oder suchen den direkten Kontakt mit uns? Let's get in touch.



Julian Schütt

Head of Business Unit
Data & Cloud Services

+49 151 64912386

j.schuett@celver.com

celver

STANDORTE

HAMBURG
celver AG
Zirkusweg 1
20359 Hamburg

MÜNCHEN
celver AG
Bürkleinstraße 10
80538 München

LISSABON
celver AG
Praça Marquês de Pombal 14
1250-162 Lisboa

KONTAKT

+49 40 468 996 4-0
kontakt@celver.com

celver.com