DOZIERENDER: MAX MUSTERMANN

# DATA ENGINEERING

THEMENLANDKARTE

| | |
|---|---|
| Datensystem-Grundlagen | 1 |
| Datenverarbeitung „at Scale“ | 2 |
| Microservices | 3 |
| Governance und Sicherheit | 4 |
| Verbreitete Cloud-Plattformen und -Dienste | 5 |
| Data Ops | 6 |

LEKTION 5

# VERBREITETE CLOUD-PLATTFORMEN UND -DIENSTE

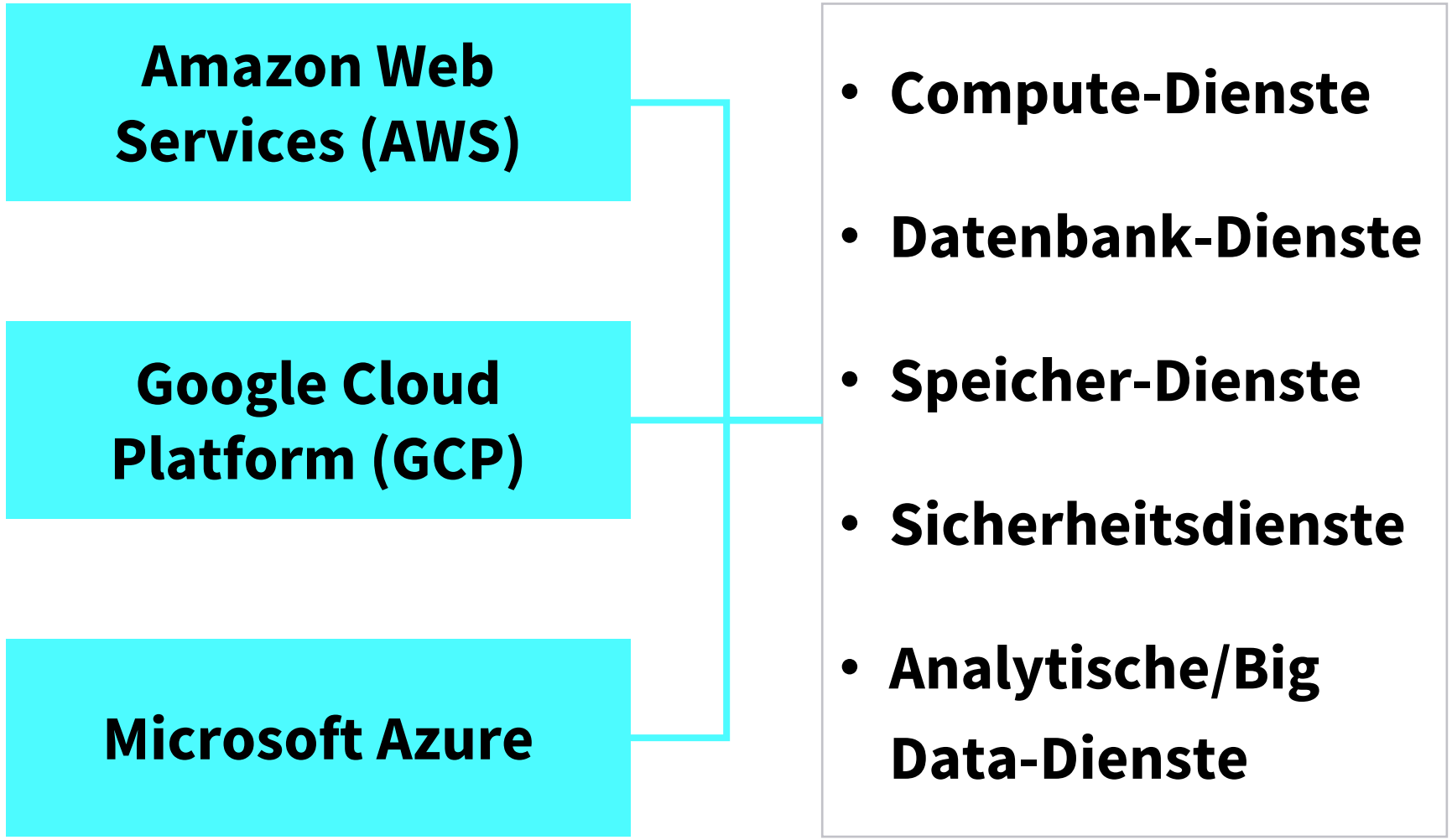
## LERNZIELE

- wiedergeben, was **Cloud Computing** ist
- erklären, was **Infrastructure-as-a-Service (IaaS)**, **Platform-as-a-Service (PaaS)** und **Software-as-a-Service (SaaS)** ist
- beschreiben, was die **Amazon Web Services (AWS)** Cloud Computing Plattform ist und welche Dienste sie bietet
- erläutern, was die **Google Cloud Platform (GCP)** ist und welche Dienste sie bietet
- beschreiben, was die **Microsoft Azure** Cloud Computing-Plattform ist und welche Dienste sie bietet

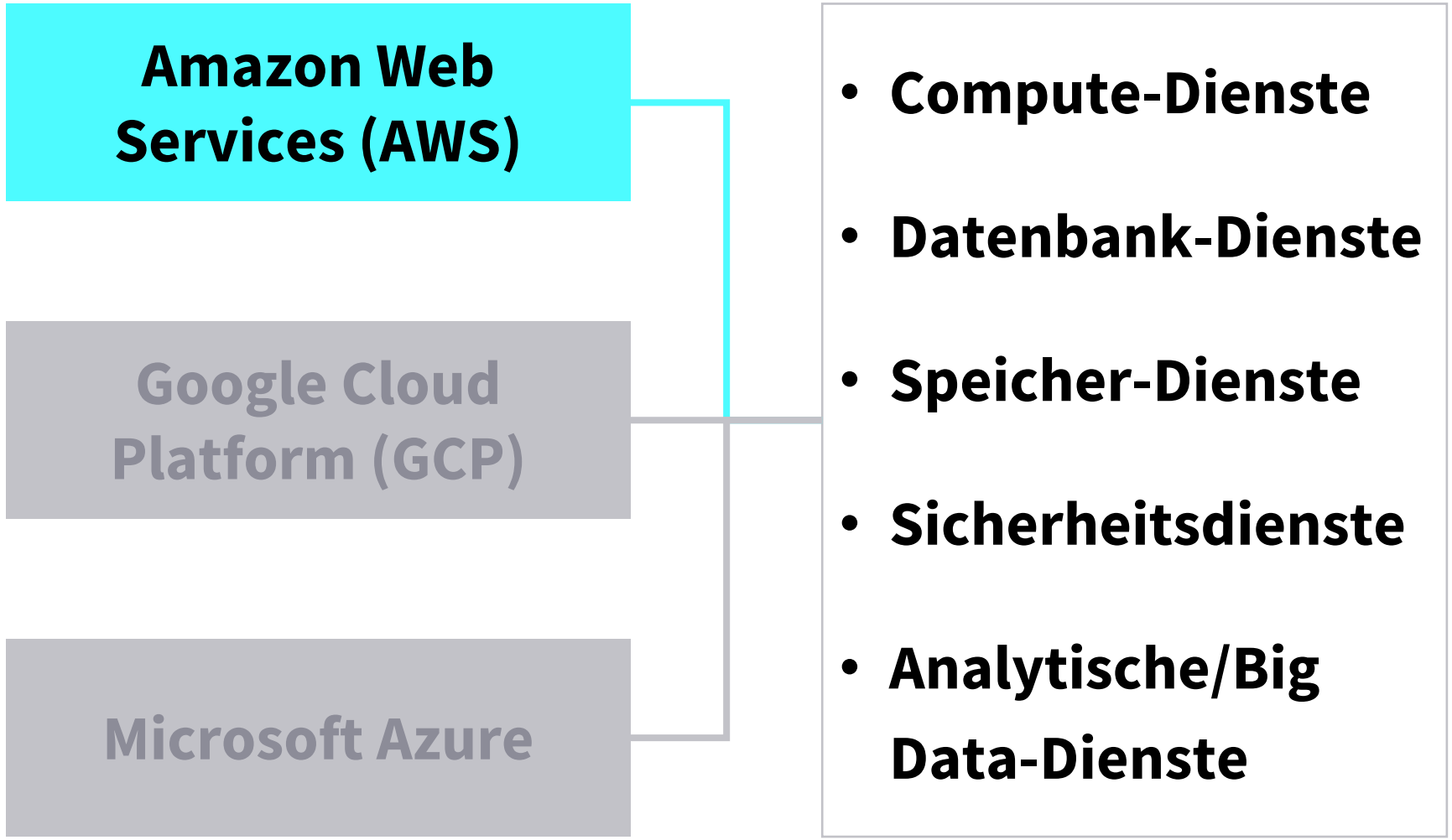
**EINFACH ERKLÄREN**

1. Beschreiben Sie die **Rechte**, die eine betroffene Person laut **DSGVO** an ihren Daten hat.
2. Erklären Sie, was der Unterschied zwischen gesicherten und sicheren Systemen ist.
3. Erklären Sie, warum es sinnvoll sein kann, ein systematisches Data Governance in einer Organisation einzuführen.

## WIEDERHOLUNG DER KERNPUNKTE DER LEKTION



## WIEDERHOLUNG DER KERNPUNKTE DER LEKTION



- Seit **2006**
- **Geographische Gliederung**
  - Region 1
    - Availabililty Zone 1
    - Availabililty Zone 2
  - Region 2
    - Availabililty Zone 3
    - Availabililty Zone 4

## Elastic Compute Cloud (EC2)

- Virtuelle Maschinen
- Preismodelle
  - On-demand
  - Reserved
  - Spot
- Wird anhand eines Amazon Machine Image (AMI) erstellt
- Nutzt Simple Storage Service (S3)
- Anzahl von VMs frei wählbar
- EC2 Auto Scaling

## Lambda

- Serverlos
- Self-contained (eigenständige kleine Programme)
- Infrastruktur vollständig durch AWS verwaltet
- getriggert durch Events

## Relational Database Service (RDS)

- Datenbank-Engine

## Aurora

- SQL und PostgreSQL

## DynamoDB

- NoSQL
- Schlüssel-Wert-Paare in Dokumenten
- hochverfügbar

## Architekturbeispiel

- Kinesis – Lambda – DynamoDB

## Simple Storage Service (S3)

- Objektspeicher-Dienst
- hochverfügbar und „haltbar“ (engl.: data durability)
- S3-Access-Point
- verschiedene Tiering Klassen (Preismodelle)

## Elastic File Sytem (EFS)

- skalierbares NFS-Dateisystem
- Linux-basiert
- verschiedene Tiering Klassen (Preismodelle)

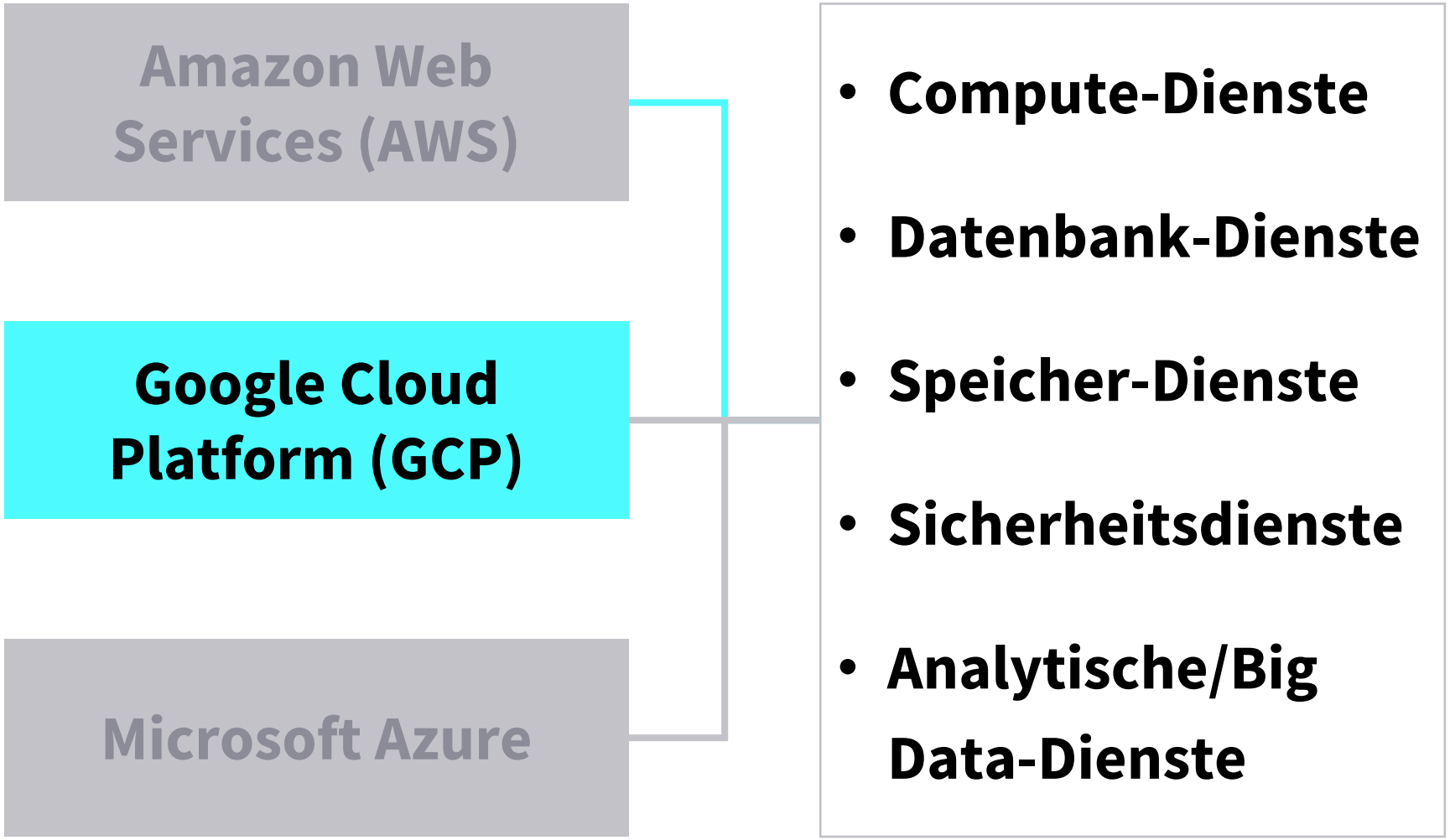
- **Security Hub**
  - zentraler Hub; unterstützt verbreitete Industrie-Sicherheitsstandards
- **GuardDuty**
  - Erkennung von Bedrohungen; maschinelles Lernen
- **Inspector**
  - Analyse von Sicherheitskonfigurationen; Vorschläge zur Verbesserung der Sicherheit
- **CloudWatch**
  - Überwachung von Protokollen und Metriken; Korrelationen und machinelles Lernen; automatisierte Aktionen

## Elastic MapReduce (EMR)

- Big Data-Prozessierung
- Verwalteter Dienst
- Basiert auf dem Hadoop Ökosystem

## SageMaker

## WIEDERHOLUNG DER KERNPUNKTE DER LEKTION



- Seit **2008**
- **App Engine** als erster Dienst
- Ressourcen werden in **Projekten** organisiert
- **Geographische Gliederung**
  - Global
    - Region 1
      - Zone a
      - Zone b
    - Region 2
      - Zone a
      - Zone b

**Cloud Functions**

- Serverlos
- Einfache Funktionen (self-contained)
- Getriggert durch Event oder Zeit-Interval
- Infrastruktur wird vollständig durch GCP verwaltet

**App Engine**

- PaaS
- Verteilte Ausführung von Applikationen
- Unterstützung vieler verschiedener Programmiersprachen

- **Google Kubernetes Engine (GKE)**
  - Containerisierte Dienste
  - Basiert auf Kubernetes
- **Compute Engine**
  - nicht durch Google verwaltet
  - IaaS
  - Volle Kontrolle über VMs

## Cloud SQL

- Vollständig verwaltete relationale Datenbank-Engine
- SQL, MySQL, PostgreSQL
- Automatisierte Backups, Replikation, Patches, Skalierung

## Firestore

- Dokumenten-basierte NoSQL Datenbank
- Schlüssel-Wert-Paar-basierte Dokumente
- in Kollektionen organisiert

## Cloud Bigtable

- Spalten-orientierte NoSQL Datenbank

## Cloud Storage

- Objektspeicher
- skalierbar
- auf große Datenvolumen ausgelegt
- verschiedene Preismodelle

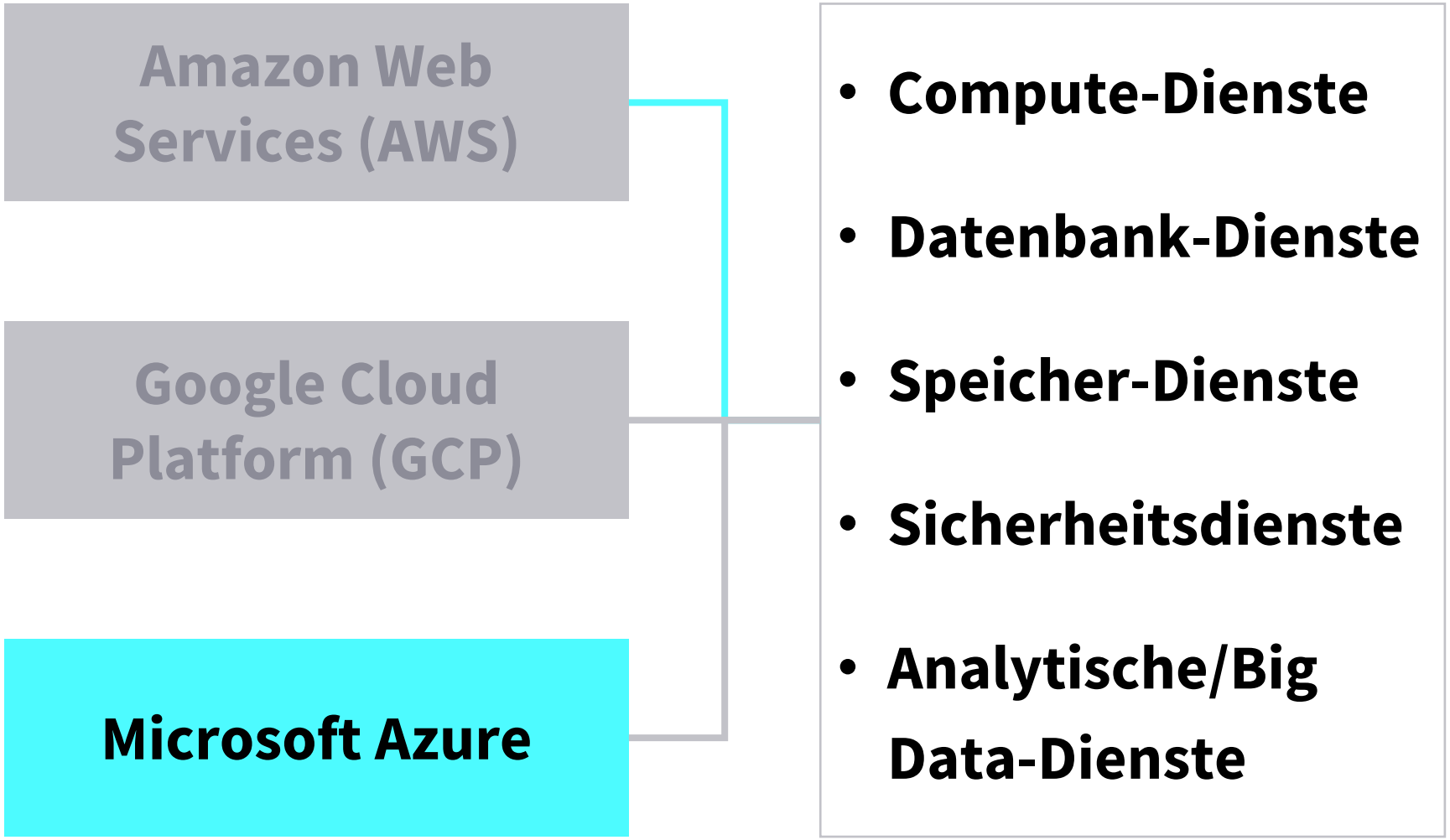
## Virtual Private Cloud (VPC)

- virtuelles Netzwerk
- Kommunikation zwischen Diensten
- globale Ressource
- Subnetzte für einzelne Rechenzentren

## GOOGLE CLOUD PLATFORM (GCP) - BIG DATA-DIENSTE

- **BigQuery**
  - Data Warehouse
  - auf sehr große Datenmengen ausgelegt
- **Dataflow**
  - Verwalteter Stream-Prozessierungsdienst
- **Pub/Sub**
  - asynchroner Nachrichtendienst
  - Entkopplung von Produzenten und Konsumenten/Abonnenten
- **Vertex AI**
  - Data Science Projekte in der Cloud

## WIEDERHOLUNG DER KERNPUNKTE DER LEKTION



- Dienste werden **Resource Groups** organisiert
- **Geographische Gliederung**
  - Global
    - Geographische Gebiete
      - Regionen
      - Gekoppelte Region-Paare
      - Rechenzentren innerhalb einer Region verbunden

## Virtual Machines

## Azure Container Instance (ACI)

- einfacher Dienst zur Ausführung von Containern zu Testzwecken

## Azure Kubernetes Service (AKS)

- verwaltetes Kubernetes Cluster

## Azure Functions

- serverlos

## Azure SQL Database

- basiert auf Microsoft SQL Server
- einzelne Datenbank
- elastischer Datenbankpool

## Cosmos DB

- NoSQL
- global
- dezentral
- Multi-Model

## Storage Accounts

- Blob
  - organisiert in Containern
  - „flat Namespace“
  - Block, Append, Page Blobs
- Files
  - Dateisystem
- Tables
  - NoSQL
- Queues
  - asynchrones Messaging

## Security Center

- Überblick-Dienst über alle Dienste

## Sentinel

- Security Information Event Management (SIEM)
- Security Orchestration Automated Response (SOAR)

## Information Protection

- Datenschutz

## Active Directory, Monitor, Advisor, ...

## Security Center for IoT

## Data Explorer

- Echtzeit-Analyse-Dienst von Datenströmen

## Databricks

- verwaltetes Spark-Cluster mit weiteren integrierten Technologien für Big Data Prozessierung

## Data Lake Analytics

## Machine Learning

## HDInsight

- wiedergeben, was **Cloud Computing** ist
- erklären, was **Infrastructure-as-a-Service (IaaS)**, **Platform-as-a-Service (PaaS)** und **Software-as-a-Service (SaaS)** ist
- beschreiben, was die **Amazon Web Services (AWS)** Cloud Computing Plattform ist und welche Dienste sie bietet
- erläutern, was die **Google Cloud Platform (GCP)** ist und welche Dienste sie bietet
- beschreiben, was die **Microsoft Azure** Cloud Computing-Plattform ist und welche Dienste sie bietet

EINHEIT 1

# TRANSFERAUFGABE

**TRANSFERAUFGABE**

Ein Start-Up das **nachhaltige Produkte in kleineren Geschäften** vertreibt war in den letzten Jahren sehr erfolgreich. In Folge sollen **weltweit weitere Filialen** eröffnet werden. Als Data Engineer:in sind Sie damit beauftragt, das **Datensystem zu entwerfen**, welches Daten über die **angebotenen Produkte** und **deren Zulieferer** speichert und verarbeitet.

Um Daten zunächst einmal im Binärformat ablegen zu können, benötigen wir einen **Storage Dienst**. Von dort aus sollen die Daten verarbeitet werden, wozu wir möglichst wenig Infrastruktur selbst verwalten wollen. Am besten wäre es, wenn wir den Code zur Datenverarbeitung hochladen könnten und uns nicht weiter um die darunterliegende Infrastruktur kümmern müssten (**serverless**).

**TRANSFERAUFGABE**

Im nächsten Schritt sollen die Daten in **streng standardisierten, strukturierten Tabellen** abgelegt werden, die miteinander schnell verknüpft und mit SQL abgefragt werden können. Parallel dazu sollten die Daten in einer **verteilten, performanten NoSQL-Datenbank** mit flexiblem Schema gespeichert werden. Von hier aus sollen mit Methoden des **maschinellen Lernens** Modelle erstellt werden, die Vorhersagen über die zukünftig benötigte Artikelmengen erlauben. Die gesamte Daten-Pipeline soll dabei durch **verwaltete Dienste abgesichert** werden.

Als Cloud-Anbieter sind wir auf die großen drei, **Amazon Web Services**, **Google Cloud Platform** und **Microsoft Azure** beschränkt. Überlegen Sie für den beschriebenen Use Case, **welche Dienste** dieser Anbieter genutzt werden können, um den Anforderungen gerecht zu werden. Zeichnen Sie für jeden der drei Anbieter ein einfaches **Architektur-Diagramm**, was den Fluss der Daten durch die einzelnen Dienste darstellt.

**TRANSFERAUFGABE**
**PRÄSENTATION DER ERGEBNISSE**

Bitte stelle deine Ergebnisse vor.
Im Plenum werden die Ergebnisse diskutiert.

1. Welcher Amazon-Dienst ist ein Überwachungsservice für DevOps-Ingenieur:innen, Entwickler:innen, Site Reliability Engineer:innen (SREs) und IT-Manager?
   a) Elastic File System
   b) CloudWatch
   c) S3
   d) Lambda

2. Wie heißt der Service von Microsoft für die Speicherung von Objektdateien, der für die Speicherung großer unstrukturierter Daten optimiert ist?
   a) Azure Files
   b) Azure Disks
   c) Azure Blob
   d) Azure Tables

3. In welchem der folgenden Cloud-Bereitstellungsmodelle haben Sie die größte Kontrolle über das System?
   a) Private (Vor-Ort-) Cloud
   b) IaaS
   c) PaaS
   d) SaaS