POLITECNICO DI TORINO

Master's Degree in Computer Engineering

Politecnico di Torino

Beyond the Hype: Strategic Cloud Infrastructure Choices for Tech Startups Balancing Innovation, Sustainability, and Resilience

Supervisors

Prof. Fulvio RISSO

Candidate

Simone ALBERTO

October 2025

Table of Contents

Chapter 1

Main Problem, Goals and Proposed Solution

Technology stack choice is one of the most fundamental and yet extensive choices that a contemporary tech startup can make. Much more than a list of individual bits of software, it is the complete master plan on which the firm is founded, and goes a long way to define a startup's operating agility, scaling, and general long-term financial sustainability. In a de facto global cloud computing infrastructure economy, these decisions become tied in with the choice of cloud service provider in a strong form of path dependency that determines the direction a company will take from its inception. [1] [2]

The overall problem solved by this thesis is the core, multi-aspect problem presented to startups by this decision-making process. The backdrop is a natural tension between accepting new, "fashionable" technologies in order to remain competitive and selecting solid, sustainable technologies that yield long-term stability and economic security. This tension has been highlighted by two dominant external forces. To start with, a more evolved and stringent regulatory framework, led by the European Union's General Data Protection Regulation (GDPR) and the EU AI Act, entails huge compliance costs, especially for startups that use AI and handle personal data. Second, rising geostrategic tensions over transatlantic data governance, as exemplified by tensions between the US CLOUD Act and EU privacy legislation, have produced deep uncertainty and led the notion of "digital sovereignty" as a technical policy challenge to a strategic business priority.

Startups are therefore in a "digital trilemma," weighing technological competitiveness, safeguarding fundamental rights, and a desire for digital sovereignty. Lacking an integrative approach, founders make early technology decisions based on short-term imperatives such as time-to-market only to be burdened with massive deferred financial expenses, refactoring spirals, and unintended regulatory hazards down the line. [28]

The primary aim of this thesis is to provide such a framework. It aims to provide startup founders, executives, and technical leaders with the strategic competencies to make well-informed, long-term, and sustainable infrastructure decisions. The particular goals are:

To break down the contemporary tech stack and perform a thorough economic examination of the underlying cloud services (compute, storage, databases, networking) offered by the three market-leading hyperscale providers: AWS, Azure, and GCP.

To explore the trade-offs of a dominant contemporary architecture—microservices on Kubernetes—as a case study for startup scalability vs. complexity vs. operational overhead trade-offs.

To supply a tangible, operational model for managing the double compliance challenge of the GDPR and the EU AI Act, correlating legal obligations to concrete cloud-native tooling and security controls.

To critically examine the case of transatlantic data flows and evaluate the strategic rationale behind the creation of a European sovereign cloud environment as a reaction to jurisdictional risks posed by non-EU providers.

For that purpose, this thesis is composed of four subsequent chapters.

Chapter 2 gives context in the form of a discussion of the strategic challenge of technology hype vs. long-term sustainability, followed by an economically driven comparison between standard cloud services and startup utilization.

Chapter 3 gives a close look at the microservices architecture orchestrated by Kubernetes and its application for startups that need extreme scalability.

Chapter 4 describes the always-critical regulatory aspect, deconstructing the GDPR and the EU AI Act and presenting a holistic framework for developing compliant AI solutions on the cloud. Lastly,

Chapter 5 interlaces these motives with a capstone examination of the geopolitical context, contending that the pursuit of digital sovereignty is an imperative strategic move and that anticipatory diversification of risks against truly European providers is a prerequisite of long-term resilience.

Finally, this thesis states that for a contemporary tech startup, its technology stack is not so much a technical realization but a strategic cornerstone. A sound choice entails a deep evaluation balancing the business model, technical infrastructure, and financial reality with the complicated legal and geopolitical terrain of the digital era.

Chapter 2

The Startup Technology Stack: A Foundation for Innovation, Sustainability, and Resilience

2.1 The Modern Startup's Architectural Blueprint

Technology stack choice is one of the most basic and essential tech startup decisions. A stack that is more than just a set of programming languages, frameworks, and software tools, the tech stack is the architectural design on which the entire company is built. This choice is one of the major reasons behind a startup's operational flexibility, its idea potency, its scalability value, and most importantly, its long-term fiscal viability and survivability. As the cloud is now the de facto infrastructure layer, the choice of cloud services provider is now no longer a standalone consideration but an integral and foundational piece of the tech stack itself.

This chapter sets the stage for the thesis's prime argument: the critical imperative for startups to navigate a balance between embracing new, occasionally "fashionable," technologies to gain competitive advantage, and choosing "stable" technologies that provide long-term stability, maintainability, and financial stability. The original choice of a technology stack creates a strong form of path dependency.

Decisions made at the beginning, driven frequently by a great push to reduce time-to-market and preserve initial investment, establish a technological and economic momentum strongly defining the trajectory of the startup in subsequent years.

This dependency can be a deep, late monetary cost. A technology chosen for its high development pace may have a short support cycle or suffer from frequent, breaking changes. For example, the AngularJS sunset caused a mid-sized SaaS firm to spend six months' worth of development effort re-writing its complete front-end, while feature development of any kind was brought to a complete standstill. Such disruptive and unintentional refactoring is proof of technical debt that is expensive in terms of deflected engineering, lost product innovation, and elevated operating risk. Thus, a strategic examination of a tech stack must look beyond its short-term

advantage to include its Total Cost of Ownership (TCO) throughout the overall product lifecycle, and this is an ongoing theme that will be examined throughout this thesis. The chapter will analyze the contemporary tech stack, examine the strategic trade-off between innovation and sustainability, and conduct a formal economic examination of the fundamental cloud infrastructure that comprises these decisions.

2.2 The Anatomy of a Startup Tech Stack

A technology stack or a solutions stack refers to the collection of technologies that an organization uses to develop and execute an application or a project. For a typical web-based startup, the stack would be divided into four simple layers:

Frontend (Client-Side): It contains all the end-user sees and interacts with on their web browser or mobile application. It's tasked with the user interface (UI) and user experience (UX). The primary technologies are HTML, CSS, and JavaScript libraries or frameworks like React, Angular, and Vue.js.

Backend (Server-Side): The heart of the application where business logic, data computations, authentications, and database queries are done. It consists of a server, an application framework, and a programming language such as Node.js, Python, Java, C#, Ruby, or PHP.

Database: This level is tasked with storing, handling, and aggregating the application data. Two broad groups of databases exist: SQL (relation), i.e., MySQL and PostgreSQL, and NoSQL (non-relation), i.e., MongoDB and DynamoDB.

Infrastructure (Cloud Services): This foundational layer supports the compute, storage, networking, and security resources upon which the application runs. For almost every startup in today's market, this tier is provided by a cloud services company such as Amazon Web Services (AWS), Microsoft Azure, or Google Cloud Platform (GCP).

From these layers, several widely-used stack archetypes have been developed with their own nature, advantages, and best-fit applications optimized for different startup paradigms.

JavaScript stacks (MERN, MEAN, MEVN): These stacks utilize JavaScript in every layer, usually supplemented by MongoDB (a NoSQL database), Express.js (a backend framework), a JavaScript frontend (React, Angular, or Vue.js), and Node.js (a server-side runtime). Cognitive ease for development teams is the greatest advantage of these stacks since one language and paradigm can be employed for frontend and backend development. Such a hybrid environment best suits to develop contemporary Single-Page Applications (SPAs) and real-time applications such as chat applications or collaboration tools.

LAMP Stacks (Traditional): One of the oldest and longest-standing open-source stacks, LAMP is made up of Linux (operating system), Apache (web server), MySQL (relational database), and PHP (programming language). Its stability, which is due to its maturity, large body of documentation, and large number of developers, and low cost make it a fine option, especially for content management systems such as WordPress and Drupal, and small and medium-sized web applications with small budgets.

Enterprise-Class Stacks (Java/Spring, ASP.NET): These stacks are known for their stability, performance, and strong security components and therefore are a top choice among large enterprise systems, financial systems, and banking systems. Java/Spring stack uses the established Java programming language with the complete Spring framework and a relational database such as PostgreSQL or MySQL. The ASP.NET stack, which is founded on the C# programming language and.NET framework, is renowned for its scalability and high performance. The newest version, ASP.NET Core, also boasts the benefit of being cross-platform with complete Windows, Linux, and macOS support.

Data-Intensive Stacks (Python/Django): This stack uses the Python programming language and the high-level web framework Django. Its biggest strength is the enormous, mature Python data science, machine learning, and AI library ecosystem. Thus, the Python/Django stack is one of the top options for data-intensive application, AI-driven platforms, and sophisticated e-commerce or content management system startups.

Serverless Stacks: The latest architecture trend conceals the underlying server management. A serverless stack will usually contain a frontend framework (such as React or Vue.js) with managed backend compute services (such as AWS Lambda or Azure Functions) and a managed NoSQL database (such as DynamoDB or Firestore). This is perfect for creating highly scalable and cost-effective applications, especially microservices architecture-based apps, event-driven systems such as the Internet of Things (IoT), as well as real-time data processing pipelines.

The decision between these archetypes is a difficult trade. The appeal of JavaScript-hub stacks, for instance, rests upon their provision of high developer velocity and cognitive ease. But this seeming ease may be concealing a substantial underlying risk: the velocity and possible volatility of the open-source JavaScript ecosystem. This is in contrast to piles such as Java/Spring, which, having a higher learning curve and initial development delay, have the advantage of having a more established, corporate-sponsored ecosystem with very predictable release timelines and long-term support promises. A startup that uses a MERN stack can probably get ahead early but silently takes on a higher chance of later refactoring churn due to

upgrades of dependencies. On the other hand, a startup opting for a Java/Spring stack embraces increased upfront complexity in return for increased long-term stability and reduced maintenance overhead—a trade-off between short-term agility and long-term sustainability.

2.3 The Strategic Dilemma: Technological Hype vs. Long-Term Horizon

The act of determining the technology stack requires the leadership of a startup to make an uncomplicated strategic trade-off: do they leverage the "hype" surrounding newest, trendiest technologies to promote innovation, or utilize a "sustainable" stack designed for long-term reliability and maintainability? Not merely a new versus old decision, but a subtle risk decision with deep financial and operational stakes.

2.4 The Allure of the New: Innovation as a Competitive Advantage

The embracing of hip, state-of-the-art technologies is usually guided by sound and reasonable business intuition. To a startup company that is trying to stand out in a saturated market space, innovation is not a luxury but an absolute need. The driving forces are mostly:

Accelerated Development: New frameworks and tools are often created with developer velocity as the design objective. In a startup that has the goal of producing an MVP as fast as possible, Ruby on Rails or those stacks based on Vue.js may be selected due to their accelerated development cycle reputation.

Talent Retention and Recruitment: The best developers tend to be attracted to projects in which they can develop new and exciting technology. With adoption of a new stack, a startup can gain a strong competitive edge in recruitment and retention of top-level engineering talent, which is valuable.

Competitive Differentiation: Emerging technologies can provide new capabilities and functionalities not found in more mature, older stacks. A company can use a new artificial intelligence platform or high-performance database to make a product faster, smarter, or less expensive than incumbents, establishing a market difference.

2.5 The Case for Sustainability: Building for Resilience and Longevity

The antithesis of pursuing technology fad is pursuing sustainability. A "sustainable" technology stack in this context does not mean an old one but a

predictable, sustainable, and long-term resilient one. The fundamental attributes of the same are driven by a strategic evaluation of its entire lifecycle.

Support Lifecycles and Predictability: A foundation of a sustainable technology is an articulated and predictable support lifecycle. This includes giving priority to technologies that get an LTS release. .NET framework, for instance, comes with LTS releases that have a guaranteed support for three years, whereas the Standard-Term Support (STS) releases are supported for 18 months only. In the long-term project, pledging an LTS version offers a secure path of patches and updates with fewer opportunities to have to do an unannounced and expensive migration. This is a choice that is really all about optimizing predictability; a reliable technology is one whose development history is clearly known and predictable, where a startup can comfortably estimate the future cost and level of effort of maintenance. The only significant financial uncertainty of "trendy" tech is the unpredictability of its future, which immediately appears as unforeseen costs of operation.

Community and Ecosystem Health: The long-term sustainability of any technology, especially open-source technology, to a large degree relies on the health of the nearby community and ecosystem. A sustainable option is one with a large, engaged community that offers excellent support, good documentation, and deep libraries and tools environment that could speed up development. An example is the Python community, which boasts an enormously vast collection of support channels in the form of official forums, Slack communities, and Discord servers, and the deepest collection of libraries for areas such as data science and AI, making it a very sustainable option for projects that incorporate those areas.

Architectural Principles and Maintainability: A sustainable stack is one that has been constructed on solid architectural principles from the beginning. That includes selecting frameworks that are modular and following practices such as Clean Architecture and SOLID design principles. These result in code that is simpler to understand, test, and maintain healthy in the long term. In addition, a sustainable architecture needs to anticipate scalability, thinking about how the system is going to expand vertically (i.e., adding power to current servers) and horizontally (i.e., adding servers). The goal is not to have a monolithic application that is going to be the growth bottleneck but rather to have a loosely coupled set of components that are upgradeable and scaleable separately.

Total Cost of Ownership (TCO): Finally, a sustainable technology solution needs to include a total cost analysis that goes beyond initial licensing or development costs. It must include the TCO, i.e., the expense of training developers in the long term, maintenance regularly, and most importantly, potential future migration or refactoring if the technology comes to the end-of-life or goes out of fashion. The case

in point here is the real usage by a bank of .NET Core 3.1 (an LTS release) that entailed a massive refactoring exercise to bring over to .NET 6 due to incompatible legacy dependencies. The problem wasn't with the technology itself, but with failing to look ahead and envision how it might unfold in the future. Sustainability is thus a conscious act of risk management, one designed to reduce the likelihood of incurring these unbudgeted expenses down the road, which is straight-up necessary for a capital-starved startup.

2.6 The Cloud as a Foundational Layer: A Strategic Comparison of AWS, Azure, and GCP for Startups

For most technology startups today, the decision is not whether to deploy in the cloud, but where to deploy in the cloud. The cloud provider is the base layer of technology stack infrastructure, and deciding between the three hyperscalers—Amazon Web Services (AWS), Microsoft Azure, and Google Cloud Platform (GCP)—is a strategic choice with broad-reaching impacts. Each of the providers has some market positioning, distinctive core strengths, and distinctive appeal to various forms of startup. [3] [4]

Amazon Web Services (AWS): Having been the incumbent market leader, AWS has the widest and most cutting-edge set of services, from basic compute and storage to sophisticated machine learning and quantum computing. Its extensive past has constructed a diverse third-party software base, a comprehensive collection of official and community-authored documentation, and the largest pool of resources of experienced cloud engineers. For a startup, this depth provides unparalleled adaptability and an abundance of resources to leverage. But the size of services itself can also come with a steep learning curve and a difficult-to-price model that is hard for small groups to master and optimize without expert advice.

Microsoft Azure: Azure has cemented its position as the good second-best in the cloud space, primarily by riding on its huge size in the enterprise software space. Its competitive strength is the capacity to integrate properly with the larger Microsoft ecosystem. For startups that develop on the ASP.NET platform, using Microsoft developer tools such as Visual Studio and GitHub, or customers that sell to enterprise clients using Microsoft technology such as Active Directory and Microsoft 365, Azure is a natural and well-integrated place to develop. This profound hybridization can meaningfully lower operational drag and lower overall cost of ownership through initiatives such as the Azure Hybrid Benefit, which allows users to leverage their on-premises Windows Server and SQL Server licenses in the cloud.

Google Cloud Platform (GCP): While a newer market entrant, GCP has forged a good reputation for technical superiority in certain, high-growth areas. It is best

known for providing the world's leading managed Kubernetes platform, the Google Kubernetes Engine (GKE), and is thus a natural fit for startups building cloud-native, containerized apps. GCP is also very strong in data analytics and AI/ML with extremely powerful and groundbreaking services such as BigQuery and Vertex AI. In addition, GCP has also emphasized heavily on a lean, developer-friendly interface and more-revealing pricing, which can simplify the learning process and lower operational overhead for committed startup businesses focused on developing API-based microservices stacks.

Selecting a cloud provider is not anymore an infrastructure choice to be made independently; it is actually a component of choosing the whole tech stack itself. The provider's managed services ecosystem, its real technical proficiency, its business model for pricing, and its financial incentive schemes become such an integral part of application development today that the provider becomes part of the backend stack. This choice has far-reaching and permanent implications for a startup's long-term TCO, efficiency of operations, and susceptibility to vendor lock-in.

Such interdependence implies the decisions on tech stack and cloud providers must be resolved simultaneously, not one after the other. For example: a startup dedicated to the ASP.NET stack 2 would discover that Azure tooling baked into it, identity services, and special licensing perks provide the development environment with significantly lower "friction cost" than trying to host the same stack on AWS or GCP. TCO on Azure would inherently be less as a result of these synergies. Likewise, a new company building a sophisticated, event-based application built from scratch to run on Kubernetes would find that GCP's more mature GKE and Cloud Run offerings are simpler to deal with and cheaper at scale than their comparable counterparts on other platforms, reducing their long-term operational expense and engineering expenses. A truly sustainable answer, then, will need to involve an in-depth analysis of the combined TCO and operating efficiency of the selected application stack for a particular cloud platform.

2.7 An Economic Analysis of Core Cloud Services for Startups

One of the most important parts of a startup's planning is a prudent economic examination of its cost of infrastructure. As the underpinning of the tech stack, the expense dynamics of the cloud provider's core offerings—compute, storage, databases, and networking—will significantly and directly determine the financial landscape of the company. This chapter offers a detailed, fact-based comparison of these offerings on AWS, Azure, and GCP considering the most appropriate pricing regimes for the life cycle of a startup. [7] [8] [9] [10]

Compute Resources: The Stack's Centerpiece

Virtual machines (VMs) are the underlying compute building blocks for the majority of cloud applications. The three big providers provide these offerings by the names Amazon Elastic Compute Cloud (EC2), Azure Virtual Machines, and Google Compute Engine (GCE). It is necessary to know their cost models in order to plan for them.

On-Demand / Pay-As-You-Go: The most versatile of the lot, this allows startups to provision and terminate compute resources without a commitment. It is billed per minute or second of usage and has the highest hourly cost. This is ideal for development, testing, and occasional workloads where flexibility matters most.

Spot Instances / Preemptible VMs:These instances provide access to an idle compute resource of a provider at 90% discounts against On-Demand pricing. The disadvantage is that the provider can terminate these instances at short notice. Therefore, they are extremely cost-effective for fault-tolerant, stateless, or batch-processing workloads of interruption tolerance but not mission-critical, stateful ones.

Commitment-Based Discounts: For predictable and steady workload startups, commitment-based discounts are the key to realizing deep cost savings. By committing to use a certain amount of capacity within a year or three years, startups can realize up to 75% savings. These are marketed as Reserved Instances (RIs) and Savings Plans on AWS 6, Reservations and Savings Plans on Azure, and Committed Use Discounts (CUDs) on GCP.

To make a material comparison, Table 2.7.1 depicts the price for a typical general-purpose VM from each of the three providers, with the cost effect of varying commitment levels.

Table 2.7.1: Comparative Pricing of General-Purpose Compute Instances

Provider	Instance Example (2 vCPU, 8 GB RAM)	On-Demand Price ($/hour)	1-Year Commitment Price ($/hour)	3-Year Commitment Price ($/hour)	% Savings vs. On-Demand (3-Year)
AWS	t3.large (Linux, US East)	$0.0832	$0.0600 (Savings Plan, No Upfront)	$0.0396 (Savings Plan, No Upfront)	52.4%
Azure	B2ms (Linux, East US)	$0.0832	$0.0570 (Savings Plan)	$0.0380 (Savings Plan)	54.3%

GCP	e2-standard-2 (Linux, us-central1)	$0.0670	$0.0422 (CUD)	$0.0302 (CUD)	55.0%

Although the discounts may appear identical, closer examination shows that there's some underlying cost of doing business associated with carrying these obligations. Choosing a discount model isn't merely an economic decision but an operational one. AWS, for example, has a very mature two-tiered model of Reserved Instances and Savings Plans. RIs offer the deepest discount for particular instance families in a particular region, while Savings Plans are more flexible at adjusting between instance types or regions. More importantly, where both are available, RIs are applied to usage before Savings Plans, introducing an additional layer of complexity needing incredibly advanced FinOps capabilities to manage most effectively and prevent wastage.

Conversely, GCP does provide Committed Use Discounts (CUDs) but then also includes Sustained Use Discounts (SUDs) to automatically apply to any resource in use for over 25% of a billing month without a commitment. This "set it and forget it" approach is very appealing to lean start-ups who might not have a full-time finance operations team. A startup may not be leveraging a complicated portfolio of AWS RIs to the best possible extent, with spending going to waste and lowering their effective discount rate. The same startup in GCP would be getting SUDs absolutely with no intervention at all. Thus, the most cost-effective model for a startup would need to factor in the Implicit Cost of Operation of running the discount strategy itself, and that might make GCP's more streamlined, automated one more suitable for resource-rationed teams.

Data Storage: The Memory of the Application

Object storage is the workhorse for storing unstructured data like images, videos, logs, and backups. The primary services are Amazon S3, Azure Blob Storage, and Google Cloud Storage. The cost of these services is multi-dimensional, depending not just on the volume of data stored but also on how that data is accessed.

A superficial comparison of the per-gigabyte storage cost is often misleading. The true cost of object storage is dictated by the application's specific I/O profile. For example, an application with a high volume of small file writes, such as an IoT sensor data logging platform, will be highly sensitive to the cost of PUT operations. In contrast, a data lake application that stores large files for infrequent, large-scale analysis will be more sensitive to the per-gigabyte storage and data retrieval costs.

This distinction is critical. For workloads that involve writing billions of small files, AWS S3 can ultimately be more cost-effective than its competitors, even if its headline per-gigabyte storage price is slightly higher, due to its comparatively lower pricing for PUT requests. Conversely, for an application storing large media assets that are accessed infrequently, the lower per-gigabyte storage cost of Azure's "Cool" tier or GCP's "Nearline" tier might be more economical. This necessitates that a startup analyze its expected data access patterns to make a financially sound choice, rather than relying on a simple price-per-gigabyte comparison.

Table 2.7.2: Object Storage Pricing Comparison (Standard/Hot Tier)

Provider	Service	Storage Cost ($/GB/month)	PUT, COPY, POST ($ per 10,000 requests)	GET, SELECT ($ per 10,000 requests)	Data Retrieval
AWS	S3 Standard	$0.023 (First 50 TB)	$0.05	$0.004	Free
Azure	Blob Storage Hot (LRS)	$0.0184	$0.055	$0.004	Free
GCP	Cloud Storage Standard (Regional)	$0.020	$0.05	$0.004	Free

Note: Prices are for US regions (e.g., AWS US East, Azure East US, GCP us-central1) and represent the first pricing tier. Prices are subject to change and vary by region and redundancy level. Azure's PUT/COPY/POST request cost is for "Write operations

Managed Databases: The Heart of the Data Layer

Managed relational databases are a cornerstone of most startup applications, which eliminate the heavy operational cost of database administration. Amazon RDS, Azure SQL Database, and Google Cloud SQL are the leaders in the market. Each of them has its own pricing philosophy and proprietary features that result in potent incentives but equally oppressive vendor lock-in.

The price and capability of these managed databases are the primary motivator behind this lock-in. For instance, Azure's Hybrid Benefit discounts companies with 40% or higher discounts if they already possess Microsoft SQL Server licenses with Software Assurance, having to pay a high monetary hurdle to escape the Microsoft universe. AWS similarly provides Aurora, its high-performance, proprietary database engine supported by MySQL and PostgreSQL, only on AWS. Its distinctive I/O-based

pricing model and tight integration with the remainder of AWS make it a highly "sticky" product. Google has joined this battle as well with AlloyDB for PostgreSQL, which it sells with strong performance guarantees and a 99.99% availability SLA, as a better quality but proprietary version of open-source PostgreSQL.

When a startup picks one of these managed databases, they are not only deciding on technology; they are also indicating commitment for the long haul to that vendor's direction. Costs of switching in data migration complexity, likely re-architecture of the application, and monetary benefits lost are high and need to be included in the first pass sustainability analysis.

Table 2.7.3 normalizes the varying different pricing plans of the three providers to allow for comparison of the cost of an average high-availability production database instance.

Table 2.7.3: Managed Relational Database Pricing Comparison (High Availability)

Provider	Service & Scenario	On-Demand Monthly Cost (Compute + Storage)	1-Year Commitment Monthly Cost	3-Year Commitment Monthly Cost
AWS	RDS for MySQL (db.t3.large, Multi-AZ, 100 GB gp3 SSD)	~$155	~$110	~$75
Azure	SQL Database (General Purpose, 2 vCore, 100 GB Storage, ZRS)	~$410	~$324	~$275
GCP	Cloud SQL for MySQL (db-n1-standard-2, HA, 100 GB SSD)	~$225	~$169	~$128

Note: This is a simplified estimation for a high-availability configuration with approximately 2 vCPUs, 8 GB RAM, and 100 GB of SSD storage in US regions. Pricing is highly dependent on the specific configuration, region, and license model (e.g., Azure Hybrid Benefit can significantly reduce Azure SQL costs). Costs are illustrative and subject to change.

Data Transfer and Networking: The Hidden Costs

Data transfer costs, particularly data egress (data transferred out to the internet), are a critical and often underestimated component of a startup's monthly cloud bill.

For any application that serves content to users over the public internet, these fees can accumulate rapidly and lead to significant budget overruns if not properly forecasted. All three major providers offer a free tier for data egress, but the pricing beyond that tier varies.

These egress fees are often referred to as a "hidden tax" of the cloud. While ingress (data transfer into the cloud) is generally free, every byte of data sent from the cloud to an end-user on the internet incurs a charge. For a startup with a successful, high-traffic application, these costs can easily grow to become a substantial portion of their total cloud spend. Table 1.4 provides a direct comparison of these crucial fees.

Table 2.7.4: Internet Data Egress Rates Comparison

Provider	Free Tier (per month)	Price per GB (First 10 TB/month)	Price per GB (10-50 TB/month)
AWS	100 GB	$0.09	$0.085
Azure	100 GB	$0.087	$0.083
GCP	100 GB (Standard Tier)	$0.12	$0.11

Note: Prices are for egress from North American or European regions to the internet and are subject to change. Pricing is tiered, and costs decrease with higher volumes. GCP offers a "Standard Tier" and a higher-performance "Premium Tier" for networking with different pricing.

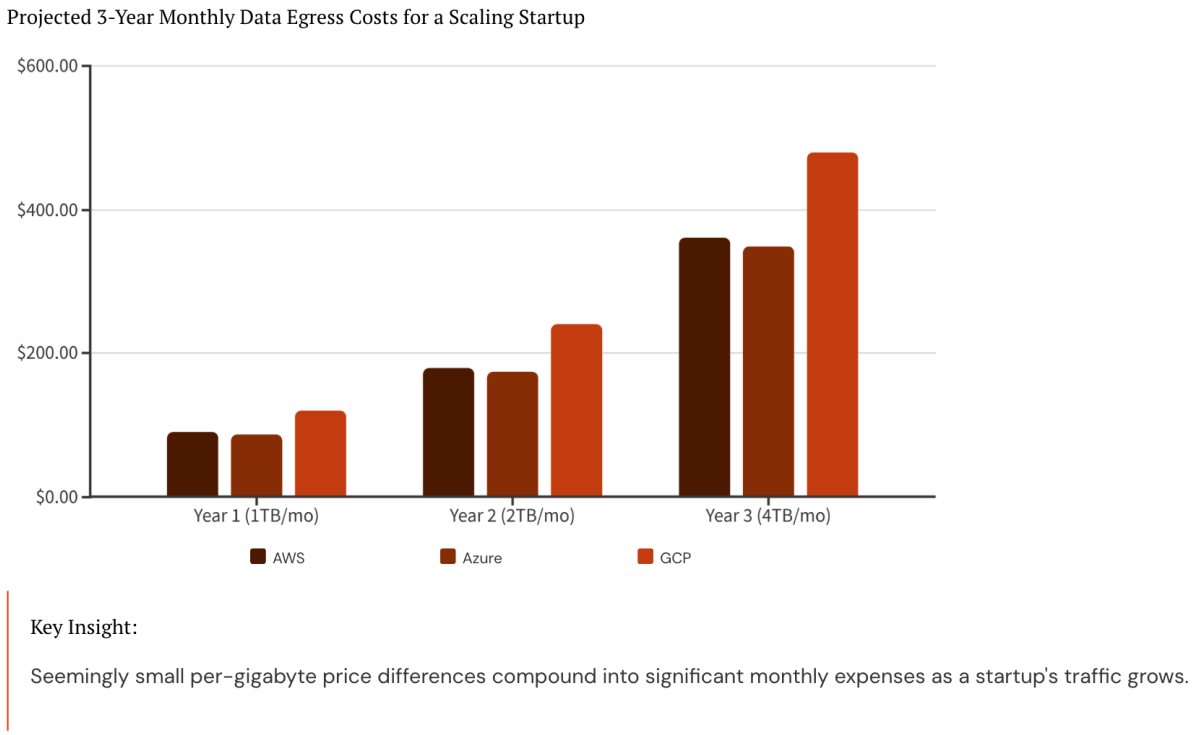


Figure 2.7.1: Projected 3-Year Monthly Data Egress Costs for a Scaling Startup

Economic Analysis of Managed Container and Serverless Platforms

Beyond traditional IaaS building blocks, startups are increasingly turning to higher-level managed platforms to accelerate development and reduce operational overhead. Managed container orchestration platforms, like Kubernetes, and serverless compute platforms have become central components of many modern tech stacks. Their economics, however, differ significantly from traditional VMs. [11] [12] [13] [14] [15]

Managed Kubernetes Platforms (EKS, AKS, GKE)

Managed Kubernetes platforms (Amazon EKS, Azure Kubernetes Service, Google Kubernetes Engine) abstract the complexity of running the Kubernetes control plane so that application runners can concentrate on running applications. Their economic designs mirror this abstraction.

Amazon Elastic Kubernetes Service (EKS): AWS levies an hourly rate for each control plane of a particular EKS cluster. Customers pay the above rate along with whatever AWS resources (for example, EC2 instances or Fargate capacity) they use to operate the Kubernetes worker nodes. Support for Kubernetes versions is also provided by AWS at a higher hourly rate with a wider maintenance window to organizations that need additional time for updating.

Azure Kubernetes Service (AKS): Azure has a tiered offering. The Free tier is zero cluster management fee and ideal for small test workloads, but still accommodates up to 1,000 nodes. The Standard tier is ideal for production and has an hourly per-cluster price for a financially-backed Service Level Agreement (SLA) and more scalability. The Premium tier has Long-Term Support (LTS) for Kubernetes versions at a higher hourly price.

Google Kubernetes Engine (GKE): GKE offers two run modes with different cost implications: Standard and Autopilot. In Standard mode, you pay an hourly management fee per cluster (with the first zonal cluster per billing account free of charge) and also charge for the Compute Engine VMs that form the cluster nodes. In Autopilot mode, however, node management is fully abstracted away. You are billed hourly per cluster (in addition to the free tier) and a charge by resources consumed (vCPU, memory, storage) by your running pods, offering a cost model nearer to actual consumption of your workload.

The selection between the two must equate the expense of the control plane itself to the expense of operating the nodes. The AKS two-tiered and GKE Autopilot mode both provide cheap entry for startups, whereas EKS is simpler but at a fixed management cost per cluster.

Table 2.7.5: Managed Kubernetes Platform Pricing Comparison

Provider	Service	Control Plane Fee	Worker Node Model	Key Differentiator
AWS	Amazon EKS	$0.10/hr (Standard), $0.60/hr (Extended Support)	Pay for EC2 instances or Fargate capacity used.	Simple but fixed pricing per cluster; extended support option.
Azure	Azure AKS	Free (Free tier), $0.10/hr (Standard tier), $0.60/hr (Premium tier with LTS) 23	Pay for the Azure VMs used. 23	No-SLA free tier ideal for experimentation; LTS option for enterprise.
GCP	Google GKE	$0.10/hr (first zonal/Autopilot cluster free per billing account) 24	Standard: Pay for Compute Engine VMs. Autopilot: Pay for resources requested by pods. 25	Autopilot mode offers a cost model based on actual workload usage, abstracting node management.

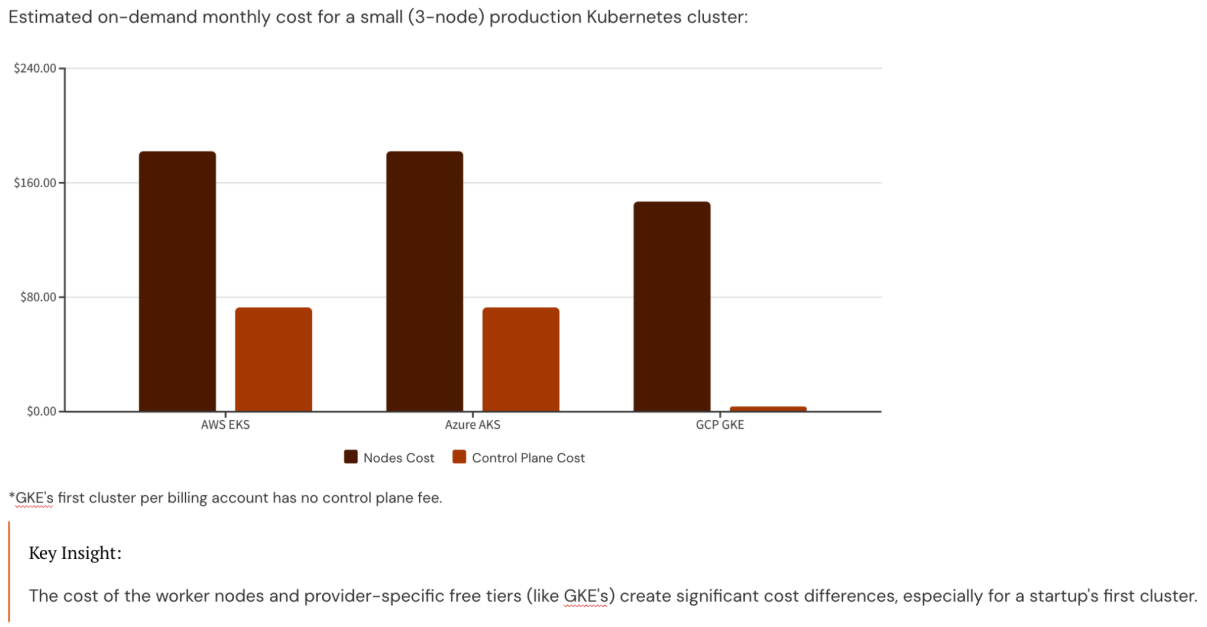


Figure 2.7.2: Estimated on-demand monthly cost for a small (3-node) production Kubernetes cluster

Serverless Container Platforms (Fargate, Cloud Run, Container Apps)

Serverless container platforms abstract further, obliterating server control in the process. Startups only get charged for the compute resources consumed when their code is executing.

AWS Fargate: Amazon ECS and EKS compatible, Fargate bills by per vCPU and memory used by the containerized workload per second with a one-minute minimum. Spiky or bursting workloads are best addressed in this model when there is an inefficiency provisioning full-time EC2 instances.

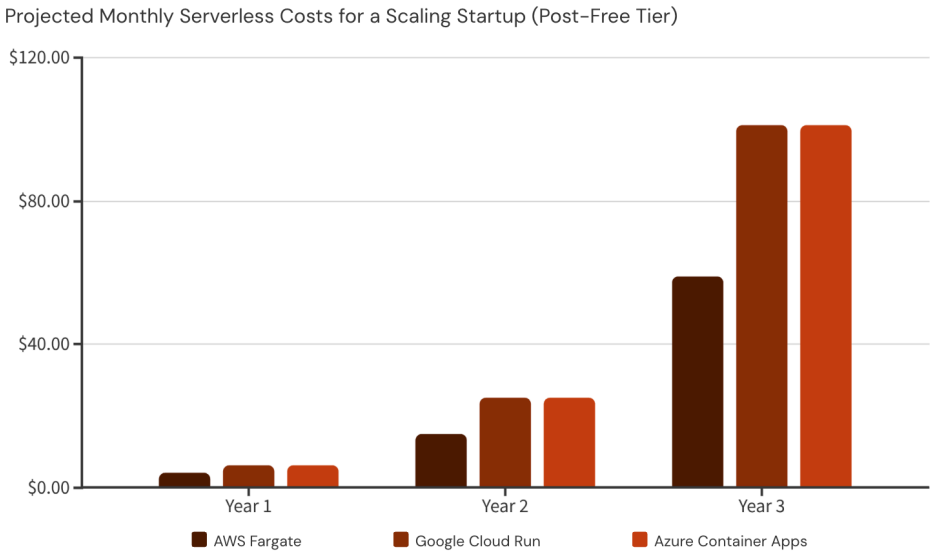
Google Cloud Run: Cloud Run is a fully managed service that bills on consumed resources (vCPU and memory) when handling requests, rounded to the nearest 100 milliseconds. It includes a very liberal monthly free tier for vCPU-seconds, GiB-seconds, and requests, and it's very affordable to get started with for startups and low or sporadic traffic applications.

Azure Container Apps: Just like Cloud Run, Container Apps also uses usage-based pricing with a monthly allowance of vCPU-seconds, GiB-seconds, and requests. Apps can scale to zero, i.e., there is no charge when the app is not running. It also allows a lower "idle" ratio for always-running instances that are not necessarily serving requests, offering a cost-against-responds trade-off.

Such platforms are well-suited to microservices architectures, APIs, and event-driven web applications, where their capability to scale down to zero can be a huge cost savings over VMs that are always running.

Table 2.7.6: Serverless Container Platform Pricing Comparison

Provider	Service	Pricing Model	Monthly Free Tier	Ideal Use Case
AWS	AWS Fargate	vCPU-hour and GB-hour, billed per second (1-minute minimum).	No Fargate-specific free tier (AWS Free Tier applies).	Running serverless containers on ECS or EKS for spiky workloads.
GCP	Google Cloud Run	vCPU-second, GiB-second, and per request.	180,000 vCPU-seconds, 360,000 GiB-seconds, 2 million requests.	APIs, microservices, and websites with low or intermittent traffic.
Azure	Azure Container Apps	vCPU-second, GiB-second, and per request.	180,000 vCPU-seconds, 360,000 GiB-seconds, 2 million requests.	Event-driven microservices and web apps with low-cost idle options.



Key Insight:

AWS Fargate can become more cost-effective at scale due to its lower per-unit compute pricing (especially on ARM architecture).

Figure 2.7.3: Projected Monthly Serverless Costs for a Scaling Startup (Post-free tier)

2.8 Financial Accelerants: Leveraging Cloud Provider Startup Programs

For pre-revenue or early-stage businesses, credits and assistance provided by the cloud providers themselves are typically the most important financial factor for the following reasons. All three players have flagship startup programs designed to discover and nurture the next generation of high-growth companies. All such programs provide a large amount of financial credits, technical assistance, and business counsel that can considerably increase a startup's runway and speed up its path to market. [16] [17] [18] [19] [20]

However, one needs to look at such programs not as charity, but as extremely effective customer acquisition and retention tools. The generous credits are designed specifically to embed a startup deeply within a provider's ecosystem, most notably inducing the use of higher-margin proprietary products. For instance, GCP's AI startup program strongly encourages the use of its Vertex AI platform 30, Microsoft's Founders Hub suggests the use of the Azure OpenAI Service, and AWS Activate suggests products such as Amazon Bedrock and its proprietary AI chips, Trainium and Inferentia.

This strategy is obviously working; statistics from Google indicate that 97% of its program startups choose to remain with GCP after exhausting their credits. The explanation is that by the time a startup is in a position to pay full subscription for its infrastructure, its architecture is well-baked at the level of the provider's specific APIs and services so that migrating to a competitor becomes excessively costly and cumbersome. Thus, startup leadership must conduct a two-fold valuation: they need to ascertain the existing value of the credits and assistance, but also the future, credit-post TCO of the respective services that they are being encouraged to embrace. Whether to take a startup package or not is really a long-term strategic decision.

Table 2.8.1 offers the multi-dimensional, structured comparison of these programs to facilitate a more strategic assessment.

Table 2.8.1: A Comparative Analysis of Major Cloud Provider Startup Programs

Program Name	AWS Activate	Microsoft for Startups Founders Hub	Google for Startups Cloud Program
Max Credits	Up to $100,000. Up to $300,000 for qualifying Generative AI startups.	Up to $150,000 in Azure credits over four years. [32]	Up to $200,000 over two years. Up to $350,000 for AI-focused startups. [33]
Eligibility	Founders Tier: Self-funded, founded < 10 years. Portfolio Tier: Affiliated with an Activate Provider (VC, accelerator), pre-Series B. [34]	Open to all startups, no funding required to apply. Tiers of benefits scale as the company grows and receives funding. [35]	Start Tier: Unfunded, founded < 5 years. Scale Tier: Funded (Pre-seed to Series A), founded < 10 years. [36]
Credit Structure	Founders: $1,000. Portfolio: Up to $100,000, typically provided in tranches based on milestones or provider relationship. [34]	Credits are provided over time as the startup grows, with different levels offering increasing amounts from $1,000 up to $150,000. [32]	Start: Up to $2,000 for one year. Scale: Up to $100,000 in Year 1 (100% coverage), plus up to an additional $100,000 in Year 2 (20% coverage). [37]
Key Non-Monetary Benefits	Technical support credits, business mentorship, access to AWS experts, exclusive partner offers (e.g., HubSpot, Deel), Startup Showcase directory. [38]	Free access to GitHub Enterprise, Microsoft 365, Dynamics 365, and Visual Studio. Technical advisory sessions, mentorship network, partner offers (e.g., Bubble, Miro). [39]	Technical support credits, access to Startup Success Managers and Customer Engineers, Google Workspace discounts, Google Maps credits, AI/Web3 specific training and resources. [40]

Note: Program details and credit amounts are subject to change. Eligibility and specific offers can vary based on the startup's affiliation with venture capital firms, accelerators, or incubators.

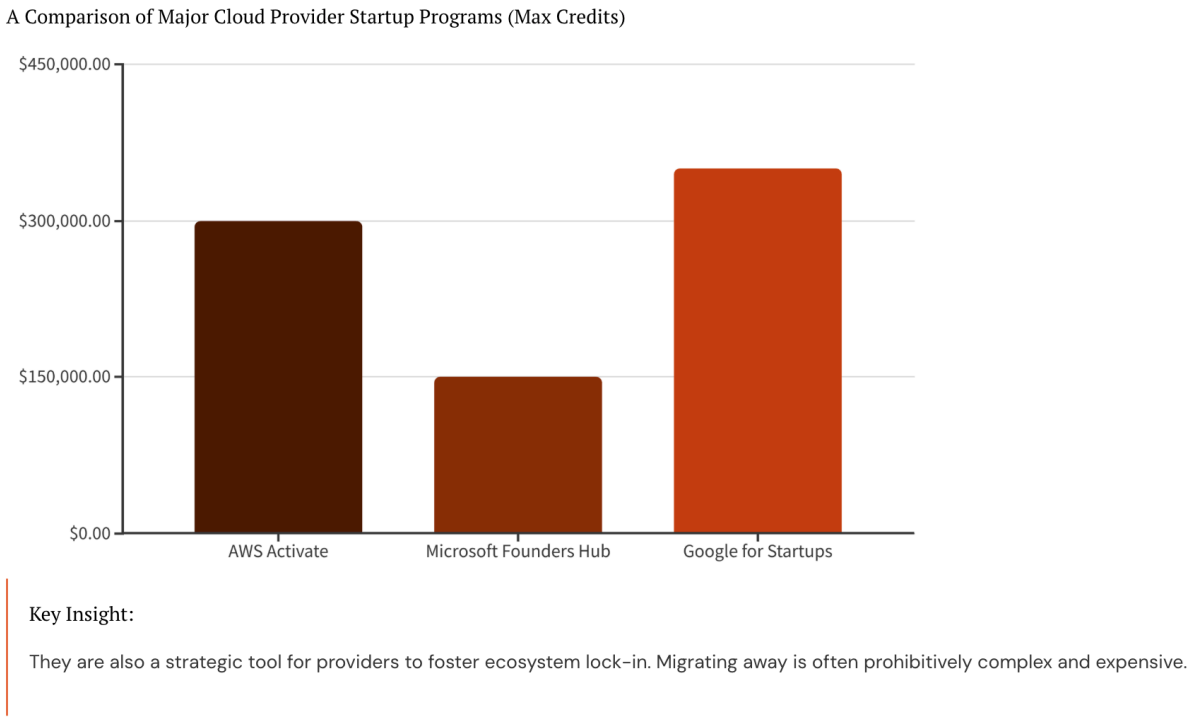


Figure 2.8.1: A comparison of Major Cloud Provider Startup Programs (max credits)

2.9 Cost Projection over a Three-Year Horizon

For a start-up, one needs to know how the cost of infrastructure varies with time in order to budget in the long term. We cannot graphically represent but can extrapolate the cost in tabular form to account for the effect of growth and the need for long-term buying strategies.

The following analysis is done assuming: a startup starts with tiny infrastructure and doubles compute capacity every year for three years. We compare month costs for On-Demand and 3-Year Commitment pricing to show potential savings.

Scenario Assumptions:

- Base Infrastructure Unit: 5 general-purpose virtual machines (roughly equivalent to AWS t3.large, Azure B2ms, GCP e2-standard-2).
- Scalability: 5 VMs Year 1; 10 VMs Year 2; 20 VMs Year 3.
- Pricing: As per numbers in Table 1.1 (US East/Central regions). On-Demand pricing is used as is. 3-Year Commitment pricing is applied to overall infrastructure as per the startup committing to its anticipated growth.

Table 2.9.1: Projected 3-Year TCO for a Scaling Compute Infrastructure (Estimated Monthly Costs)

Year	Resources (VMs)	Provider	Estimated On-Demand Monthly Cost	Estimated 3-Year Commitment Monthly Cost	Estimated Monthly Savings
Year 1	5	AWS	$303	$144	$159 (52%)
		Azure	$303	$139	$164 (54%)
		GCP	$245	$110	$135 (55%)
Year 2	10	AWS	$607	$289	$318 (52%)
		Azure	$607	$277	$330 (54%)
		GCP	$489	$220	$269 (55%)
Year 3	20	AWS	$1,215	$578	$637 (52%)
		Azure	$1,215	$555	$660 (54%)
		GCP	$978	$441	$537 (55%)

Note: Projections are based on 730 hours per month and the per-hour prices from Table 2.9.1. This is a simplified model and does not include storage, networking, or other service costs. The purpose is to illustrate the financial impact of scalability and commitment discounts.

This tabular projection clearly demonstrates a fundamental principle for a startup's financial sustainability: as infrastructure grows, the impact of long-term commitment discounts becomes exponentially more significant. In Year 3, a startup on GCP could save over $500 per month by committing to a three-year plan versus paying on-demand. For a startup looking to maximize its runway, the ability to forecast workloads and leverage these discounts is not just a cost optimization, but a strategic necessity for long-term survival and growth.

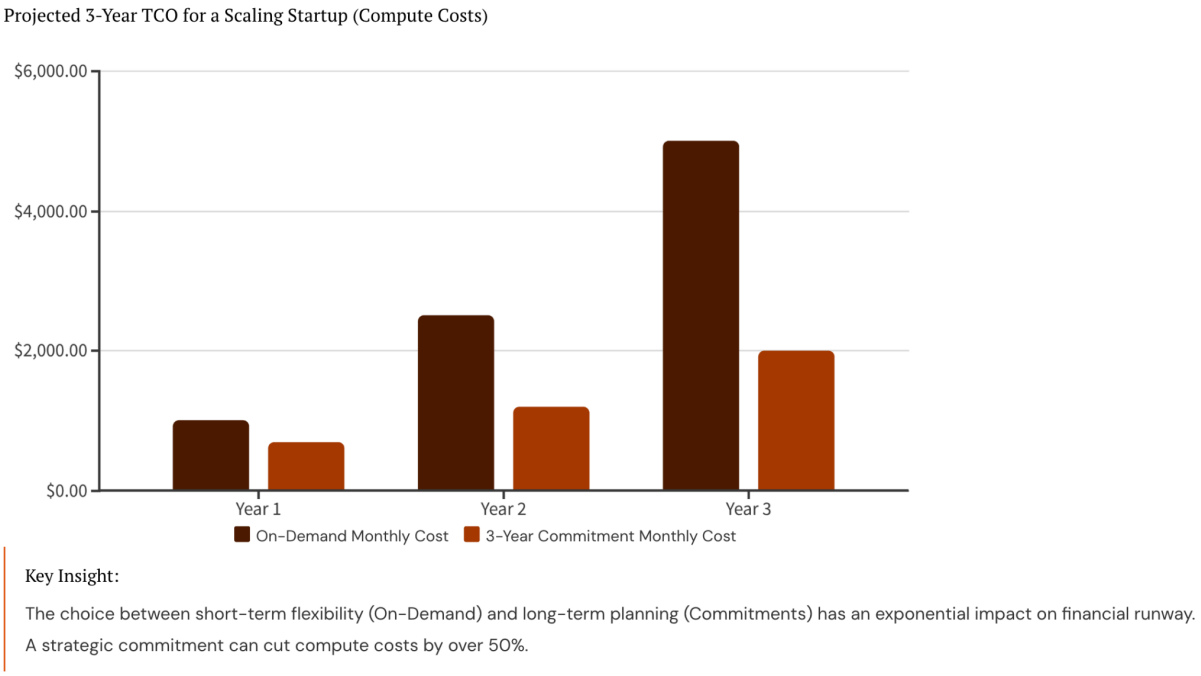


Figure 2.9.1: Projected 3-Year TCO for a Scaling Startup (compute costs)

2.10 Synthesizing a Strategic Stack Decision

The technology stack choice is a formative moment in the life of a technology startup, an engineering choice whose implications reach far beyond the engineering team deep into the financial and strategic heart of the business. In this chapter, it was demonstrated that the choice is not merely choosing the hippest or "best" technology but a careful balancing act between the cross-cutting demands of innovation, financial viability, and operational longevity.

The analysis has found a chain of these critical trade-offs which all start-ups must confront. There is the trade-off between short-term speed of development offered by hip, new frameworks and long-term maintenance cost and migration risks posed by their rapid evolution and shorter support lifecycles. There is the trade-off between cognitive simplicity of a single language stack and predictability and stability of an established, enterprise-class technology. There is a cost trade-off between the service top-line price and its actual Total Cost of Ownership, which has to factor in the application's particular usage characteristics and the implied cost of operations of having to deal with sophisticated discount structures. And then there is the strategic trade-off among the huge upfront worth of cloud credits from early startup initiatives and the long-term vendor lock-in threat that such initiatives carry.

Lastly, the winning and long-term technology stack will not be a product of some deus ex machina that will somehow "pick" in mid-air. Instead, it will be in ideal

harmony with the startup's distinct circumstances. This involves a comprehensive analysis that links the company's target market and business model to its technical foundation, and locates both within the fiscal realities of its capital schedule and longer-term cost outlooks. The most enduring startups will be the ones that are able to look beyond the current hype, and view their technology stack not as a collection of trendy tools, but as the strategic, underlying asset on which they will establish their long-term success.

This long-term TCO analysis is not merely theoretical. As the cloud market matures, a counter-trend of 'cloud repatriation' is emerging, driven by the very cost, performance, and control issues that a sustainable stack selection aims to mitigate. An analysis of this trend, exemplified by high-profile companies, provides valuable lessons for startups planning their long-term infrastructure strategy.

2.11 The Great Recalibration: Why Some Tech Giants Are Moving Back from the CloudFor over a decade, the "cloud-first" mantra has dominated IT strategy.

The promise of infinite scalability, ease to run, and cost savings through pay-as-you-go fueled a gargantuan shift to public cloud infrastructures. But as we move into 2024 to 2025, the opposite narrative starts: cloud repatriation. This is not an exodus, but a deliberate rebalancing, in which experienced organizations are deliberately moving workloads off the public cloud to on-premises or hybrid environments.

This "cloud reset" is the result of the convergence of issues to the original, often euphoric, vision for the cloud. Surveys toward the end of 2024 point to a robust upward movement, with one study by Barclays discovering that a record 86% of CIOs intend to shift at least some workloads off the public cloud.

Though total spending on cloud is still expected to increase, the repatriation trend underscores a maturing market in which organizations are shifting away from a "cloud-first" towards a "cloud-smart" strategy and looking for a balance of cost, control, and performance.

The Primary Drivers of Cloud Repatriation

The reversal from the public cloud is rarely because of some isolated issue in and of itself. Instead, it is a tactical response to an extended series of related issues that compound as the cloud infrastructure of an organization grows.

Spiraling and Unpredictable Costs: It is the most widely cited reason. The initial glamour of trading run-rate operating expenses of real estate for trading capital

investment in real estate usually gives way to "bill shock." As companies expand, pay-as-you-go can be "grotesquely expensive," especially on predictable and reliable workloads. Sneaky charges, dazzling pricing models, and notably steep data egress fees (the cost of moving data out of the cloud) can result in runaway opex far exceeding initial cost savings. The Flexera 2025 State of the Cloud Report finds companies are estimating up to 27% of cloud spending is wasted.

Performance and Latency Issues: for particular mission-critical workloads, public cloud infrastructure latency isn't acceptable. Companies are finding that on-prem or at the edge locations can deliver better performance and more predictable response times by placing workloads near the point of consumption. That's particularly true for HPC and deep AI model training workloads.

Security, Compliance, Data Sovereignty: with global data privacy legislation like GDPR still in force, the importance of total control over data is priority number one. Having sensitive data in a multi-tenant public cloud raises sticky issues over data residency and ownership of access. Repatriation allows companies to have tailored security controls and meet strict data sovereignty regulations that mandate data stay within geographic limits.

Vendor Lock-In: most organizations become deeply embedded in the tech stack of a single vendor, reliant on proprietary services and APIs. This makes it technologically advanced and prohibitively expensive to change vendors or move on-premises, cutting strategic flexibility and bargaining power.

Case Studies in Cloud Repatriation

Several high-profile companies have publicly detailed their journey back from the cloud, providing concrete examples of the motivations and outcomes of repatriation.

1. 37signals (Basecamp & HEY)

Arguably the loudest cloud exit proponent, software firm 37signals started moving its products, including cloud-born email service HEY, back to AWS and Google Cloud towards the end of 2022. Price was a key motivator. Co-founder and Chief Technology Officer David Heinemeier Hansson labeled their $3.2 million annually cloud cost "obscene" for a mid-sized firm that has average workloads. He wasn't convinced leasing computers from cloud vendors is a bargain for established organizations that can split the hardware's cost over three or four years. Process & Results: They spent around $700,000 on their own Dell servers, which were kept in a colocation center. As of October 2024, Hansson reported their yearly cloud cost had declined to $1.3 million from $3.2 million, down by nearly $2 million a year. Its initial hardware investments were completely covered by the savings realized in just half of 2023 alone. With a plan to migrate their remaining 10 petabytes away from Amazon

S3 later this summer of 2025, 37signals conservatively estimates total five-year cost savings of over $10 million. [21]

2. GEICO

The Berkshire Hathaway-owned insurance giant represents a large-scale enterprise example of cloud repatriation. After a decade-long journey into the cloud that began in 2013, GEICO is now undertaking a massive infrastructure overhaul to bring many workloads back on-premises. Reason: A combination of spiraling costs, declining availability, and the need for greater control over data for compliance and AI initiatives. By 2021, GEICO was investing over $300 million annually on eight different cloud providers and had 80% of its workloads running in the public cloud. All that bulky, multi-cloud configuration created more reliability problems and an overall data strategy absence. Process & Outcomes: GEICO began building a massive private cloud platform on Open Compute Project (OCP) equipment in 2023. The transition already has realized incredible returns: a 50% reduction in computer expense and a 60% reduction in storage expense. The company is pulling out all stops to hire to ramp up on-premises capacity, evidence of long-term strategic commitment to hybridization. [22]

3. Ahrefs

Singapore-headquartered search engine optimization solutions provider made a strategic choice to avoid a massive cloud migration outright, choosing instead to invest heavily in its own on-premises infrastructure from the beginning. Why: A preliminary cost versus benefit study revealed that a strictly cloud-based solution would be economically unviable. The company estimated it would cost over 10 times as much to put its infrastructure up on AWS compared to placing its hardware in a colocation facility. Process & Results: Ahrefs has invested $122 million in on-prem infrastructure between 2017. It approximates that the same setup on AWS would have cost more than $1.1 billion over the same time period. Not only did this on-prem solution save the company hundreds of millions of dollars but also gave better performance with quicker, more powerful servers than comparable cloud instances. [23]

4. Dropbox

Among the earliest and most prominent, file-hosting company Dropbox launched its "Infrastructure Optimization" initiative in 2015, relocating much of its data away from AWS onto its own custom infrastructure. Why: As the company expanded and prepared for an IPO, pressure to expand profit margins and assume more control of its core storage infrastructure became critical. On that metric, developing in-house was more economical. Process & Results: Dropbox migrated about 90% of its users' data to Dropbox's internal colocation data centers. The migration had a $75 million

impact on operating costs during the first two years. Worth noting, however, is that Dropbox did not completely move away from the cloud; it still uses AWS for worldwide access, especially in Europe and Asia, showing the strategic merit of a hybrid strategy. [24]

Chapter 3

Microservices Architecture on Kubernetes for a tech Startup

3.1 Fundamentals of Microservices

The design of microservices is a major step forward in the development of software from monolithic programs to the distributed form where an application is organized as a set of extremely small, autonomous services. They talk to each other on the network utilizing usually very lightweight protocols. A single microservice handles a specific business capability, and thus there can be high modularity as well as separation of concerns. This pattern is different from monolithic applications where all is tightly coupled into one codebase. The movement towards microservices is attributed to the rise in complexity and scalability demands of current applications. With more and more applications, monolithic frameworks tend to be painful to scale, maintain, and update effectively. [25]

Most important features of the microservices approach:

- A service is a deployable unit: provides a clearly defined client interface (end-points, methods, data types, relations, …) and satisfies agreed Service Level Agreement (availability, scalability, resilience, …).
- A service has its own database where it stores its own information and duplicates other services' information, if necessary.
- A service can be a consumer of services and can be mapped into a common communication bus where a suitable event is published and subscribed.

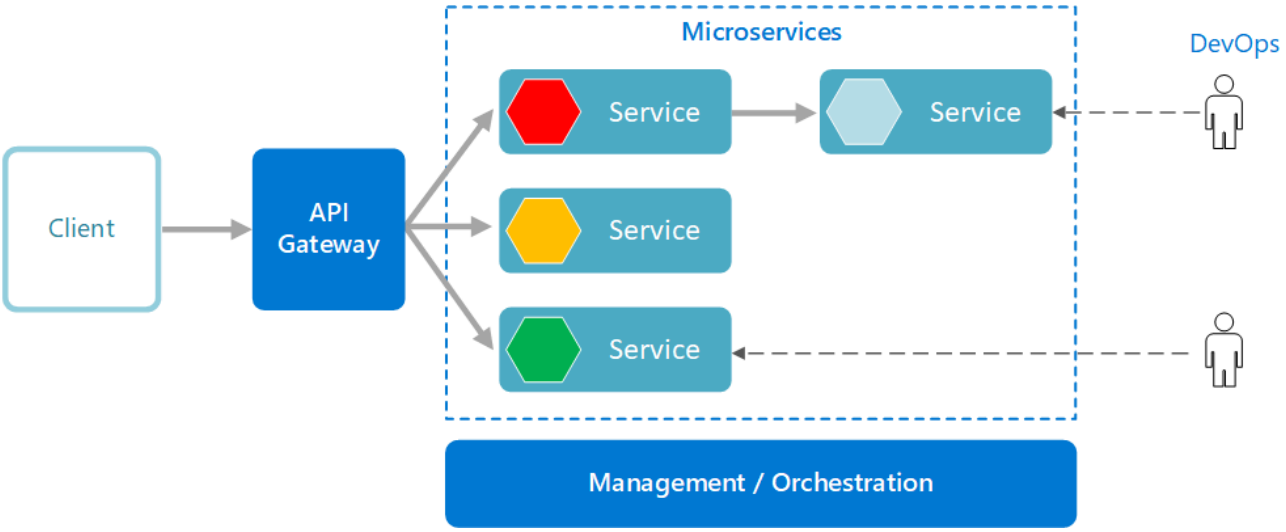


Figure 3.1: High level vision of a microservice architecture

Service features

An effectively designed microservice has some key qualities that make a microservices architecture fault-resilient and agile in general:

- The most testable and easiest to maintain is first, which leads to rapid development iterations and efficient deployment. This is made possible because there is appropriate separation of concerns and an emphasis on modularity.
- Loosely coupled with other services so that teams can work in isolation and the effect of changes is minimized across the system.
- Independently deployable: allowing teams to publish updates without having to coordinate with other teams.
- Able to be developed by a small team: needed for optimal productivity, avoiding communication overhead of big teams and creating a more concentrated and productive development process.

3.2 Kubernetes Overview

Kubernetes has become the de facto standard for application deployment, scaling, and automating management of containerized applications. It offers a very solid foundation for container orchestration, collecting them in logical groups (Pods) to make managing and locating applications straightforward. The development of Kubernetes is inherent to the development of application deployment strategy and exhibits an ongoing pursuit of efficiency and scalability.

An Historical Perspective on Deployment Evolution

Kubernetes' evolution started from the Traditional Deployment Era, where applications were deployed directly on bare metal servers. This method was straightforward but low in resource utilization and scaling. The Virtualized Deployment Era came with virtual machines (VMs), where many applications could share a single physical server. This was very optimized in resource utilization but still had the overhead of operating system virtualization. The Container Deployment Age changed application deployment in the form of light containers that share a common operating system host, yielding improved portability, resource utilization, and environmental consistency. [26] [27]

Kubernetes: A Platform for Resilient Distributed Systems:

Kubernetes was developed to overcome the issues of running containerized applications in production. It offers an end-to-end system for constructing fault-tolerant distributed systems, and automated operations like scaling, failover, and service discovery. Some of its core features that showcase its capability are:

- Service Discovery and Load Balancing: Kubernetes offers easy discovery of services by using DNS or IP addresses and routes the network traffic in a way so that the application is stable under different loads.
- Storage Orchestration: It automates storage system provisioning and mounting, masking the underlying storage infrastructure.
- Rollouts and Rollbacks: Kubernetes facilitates declarative deployment management, ensuring safe updates and rollbacks to preserve application integrity.
- Self-Healing Capabilities: It constantly monitors app health, automatically restarting failed containers, replacing hung ones, and accepting traffic only from healthy instances.
- Secret and Configuration Management: Kubernetes offers safe storage and management of sensitive information, improving the security of the applications.
- Horizontal Scaling: It makes scaling applications easy, enabling dynamic scaling based on need or utilization.
- Extensibility: Its modular architecture enables feature extension and customization without altering core components.
- Dual-Stack IPv4/IPv6 Support: Kubernetes natively supports both IPv4 and IPv6, and hence it is compatible with multinet environments.

Internal Architecture Overview

A Kubernetes cluster consists of machines, or nodes, that run containerized applications. Behind these lie some underlying concepts. A Pod is the most basic deployable in Kubernetes and is a group of one or more containers that share a common underlying network and storage resource. Nodes are worker machines upon which Pods are executed. A Cluster contains one or more nodes controlled by a control plane. Namespaces are utilized to partition resources logically in a cluster. Deployments are an abstract top level, which control the creation and life of Pods in such a manner that there are a set number of replicas at any given time. Services provide a stable DNS name and IP address by which the Pods are accessible so that service discovery and load balancing are possible. Ingress enables management of access from outside the cluster into cluster resources, often HTTP or HTTPS.

Kubernetes topology is on top of a control plane and worker nodes. The control plane manages the cluster and has items such as the API server (the Kubernetes control plane's front-end), etcd (cluster configuration data storage, distributed key-value database), scheduler (schedules Pods onto nodes), controller manager (executes multiple controller processes which manage cluster state), and cloud-controller-manager (cloud provider bridges). The worker nodes run the actual application and store things like kubelet (a running process on each node that is in charge of the Pods), kube-proxy (a network proxy on each node that is in charge of the network routing), and a container runtime (like Docker or containerd) that runs the containers.

Kubernetes provides an abstraction layer above infrastructure to enable application developers to focus on application logic rather than on container or virtual machine life cycle intricacies. Kubernetes adds operational complexity to distributed application runtime through automated scaling and self-healing mechanisms. Also, declarative Kubernetes configuration in terms of what the system should be and not how to get there allows for reproducibility and consistency of deployments, thus reducing the likelihood of the system moving away from its configuration at runtime.

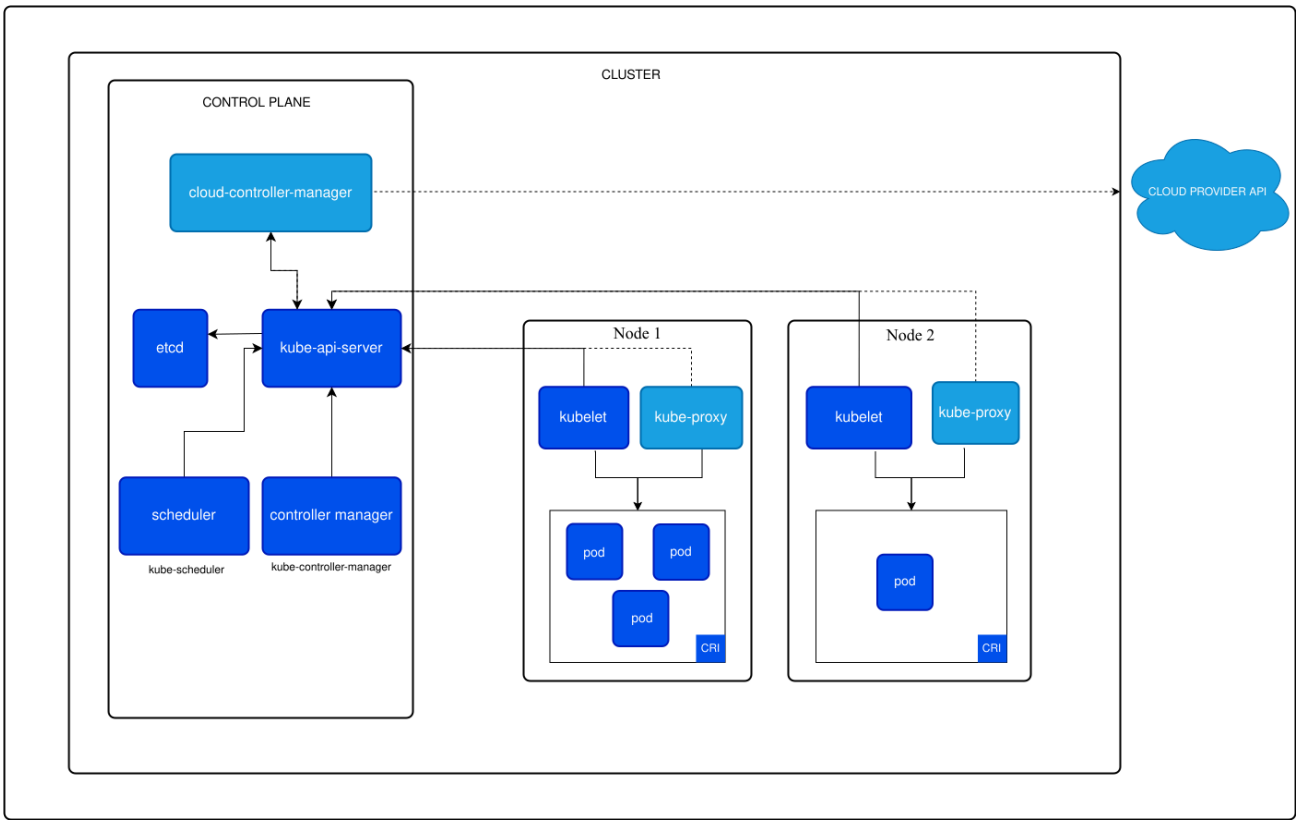


Figure 3.2: Kubernetes cluster components

The Synergy Between Microservices and Kubernetes

Kubernetes is generally accepted as a natural and natural environment for microservices deployment and management. Most of the problems of running a distributed system, comprising many independent services, are automatically managed by the platform.

Certainly, the biggest synergy is the way in which service discovery is enabled in Kubernetes. Where microservices architecture is present, the services must have the ability to find one another and exchange information. Kubernetes has built-in service discovery, in the sense that services can discover one another by using DNS names or environment variables. Load balancing is also a significant operation in the management of microservices, as it distributes incoming traffic to numerous instances of the same service. Kubernetes Services automatically offer load balancing, which enhances application availability and performance.

Scalability

Scalability is one of the fundamental requirements of most microservices architectures, and in this regard, Kubernetes excels. It allows for horizontal scaling of each microservice as needed, and dynamically adjusts the number of instances running to match the workload. This ensures maximum resource utilization and allowance for peak loads. Fault tolerance is also critical within a distributed system. Kubernetes is always monitoring the health of Pods and can automatically restart a crashed container or replace a dying node, hence adding to the overall stability of the application. Microservices and Kubernetes combined make a highly dynamic and fault-tolerant system in which any service can be independently scaled depending on demands, resulting in maximum use of resources. Furthermore, the self-healing of Kubernetes, i.e., restarting crashed containers automatically and replacing faulty nodes, also improves the overall reliability and availability of the platform.

3.3 Analyzing the Advantages and Disadvantages of Kubernetes for a Startup

The choice by a startup to use Kubernetes is a representative microcosm of the strategic trade-off between technological novelty and long-term sustainability discussed in Chapter 2. Although the platform can provide enormous scalability and is the contemporary default for cloud-native deployment (the "charm of the new"), its inherent complexity can create enormous risks to the financial and operational sustainability of a startup.

Advantages

Kubernetes has a number of significant benefits to a startup. One of the biggest benefits is scalability because Kubernetes enables pulling and scaling individual microservices dynamically to suit changing workloads. This facilitates the platform to grow easily in response to growing user demand and transaction volume without making huge architectural modifications. For a startup which is expected to scale rapidly, Kubernetes offers an infrastructure that is future-proof and can scale elastically to keep up with increasing user demand and data without recourse to the level of basic architectural changes. Scale driven by demand will be what sustains a startup, and Kubernetes has the building blocks to enable this scaling.

Resource usage is another important benefit. Kubernetes maximizes the utilization of compute resources by containerizing in an optimal manner and loading them onto available nodes. This can result in massive long-term cost savings.

Fault isolation is further improved with Kubernetes because failures within a microservice will not readily spill over into other areas of the application and hence render the system less stable overall.

The healthy and wealthy ecosystem support surrounding Kubernetes is priceless. The platform has a mammoth developers' and operators' base working on it, and plenty of tools and extensions exist that can be used to further its capabilities. With the large and established community of Kubernetes, there is plenty of knowledge, support, and easily available solutions to shared issues awaiting, lessening the learning curve and possible hurdles to the startup's development team. Strong community offers rich resources, and issues therefore become simpler to repair and best practices simpler to remain up-to-date about.

Disadvantages

Even though it has many benefits, the application of Kubernetes also has some disadvantages, especially to a startup.

The inbuilt complexity of Kubernetes will often be a major obstacle. It is a high-powered platform with many concepts and things to know, so it has an extremely high learning curve for dev and ops teams. This requires time and capital to train and reskill.

Although Kubernetes has a massive long-term payoff, its initial complexity and learning curve may be a hurdle to a small business with a limited budget and personnel. Startups will have tight budgets and small personnel lists, and investment in learning and deploying Kubernetes upfront may be substantial.

This high learning curve and operational sophistication directly contributes to a greater Total Cost of Ownership (TCO), precluding the simple on-demand models discussed in Chapter 2. A startup company must consider the high, sometimes opaque, expense of employing or training knowledgeable DevOps engineers, who command premium salary rates. This can significantly skyrocket operating costs, directly threatening long-term financial sustainability needed to survive.

The initial cost of installation in terms of effort and time to deploy and configure a Kubernetes cluster can be substantial. Additionally, there is a recurring operational expense of running and hosting a Kubernetes cluster, e.g., monitoring, upgrading, and debugging. A thoughtful plan of the startup team size and possibly usage of managed Kubernetes platforms (e.g., AWS EKS, Google GKE, Azure AKS) can reduce the operational cost and overhead of self-hosting a Kubernetes cluster. Managed services reduce the complexity of Kubernetes management and thus make it more appealing for startups.

Although the managed Kubernetes offerings, such as AWS EKS, Google GKE, and Azure AKS, are constantly marketed as the solution to complexity, they pose an added strategic risk: vendor lock-in. By creating applications that are dependent on a provider's own Kubernetes integrations, custom extensions, and APIs, a startup weakens its potential to migrate. This reliance complicates the possibility of achieving a multi-cloud approach to diversify risks, another of the principal proposals that arise from the discussion of geopolitical and jurisdictional risks in Chapter 5.

In addition, from a regulatory perspective, the complexity of the platform is also a major security concern. A minor misconfiguration of the complex network policies or access controls within Kubernetes can leave sensitive information exposed. An incident like this would have transparent implications for a company's compliance with Article 32 of the GDPR ('Security of Processing') and with the cybersecurity resilience obligations of high-risk AI systems under the EU AI Act.

Table 3.3.1: Advantages and disadvantages

Feature	Advantage/Disadvantage	Relevance to Startup
Scalability	Advantage	High
Resource Efficiency	Advantage	Medium
Fault Isolation	Advantage	High
Ecosystem Support	Advantage	High
Complexity	Disadvantage	High
Learning Curve	Disadvantage	High
Initial Setup Costs	Disadvantage	Medium
Operational Overhead	Disadvantage	Medium

Chapter 4

Navigating the New Regulatory Gauntlet: A Framework for Security and Privacy in Cloud-Based AI Systems under GDPR and the EU AI Act

4.1 Foundational Principles of Data Protection and AI Governance

The development and deployment of Artificial Intelligence (AI) systems, specifically those built and deployed on cloud-based scalable infrastructure, have brought with them unprecedented amounts of data processing and automated decision-making. This technological development has also created a new and challenging regulatory environment, led by the European Union. Two pillars of legislation today set the parameters of responsible innovation: the General Data Protection Regulation (GDPR) that sets a uniform framework for protection of personal data, and the historic AI Act, which adopts a risk-based approach to regulation of AI systems themselves. Familiarity with the different yet complementary principles of these two regulations is a doorway to all organizations that wish to engage AI solutions in the European market. This chapter will set the legal foundation on which the following technical analysis is constructed, deconstructing the fundamental principles of both GDPR and the AI Act to build an understandable, legally accurate view of the world of compliance.

4.2 The General Data Protection Regulation (GDPR): A Data-Centric Paradigm

The General Data Protection Regulation (Regulation (EU) 2016/679) is a paradigm shift in data protection law, putting in place a harmonized framework of rules across the EU and bestowing individuals with great rights over their personal data. Although its reach is extensive, applying it to AI systems is challenging because of

the scale, intricacy, and obscurity of data processing involved in machine learning. [29]

Core Principles

The GDPR is based upon a framework of principles underlying any processing of personal data that will need to guide all such processing. These are the transparency, fairness, and lawfulness principles requiring data subjects to be informed in an understandable way of how their data is being processed. The purpose limitation and data minimisation principles require data to be collected for definite, specified purposes and only that data which is required for such a purpose to be processed. The accuracy principle requires personal data to be current and correct. These types of principles are typically put through scrutiny by AI models that would handle massive volumes of information for uses that differ throughout model creation and might set up new, possibly erroneous, information regarding individuals.

Data Protection by Design and by Default (Article 25)

Article 25 of the GDPR is a bedrock of its regulation of AI, in which data protection needs to be integrated into the design of processing systems from the very beginning. It is not an afterthought procedural requirement but rather a design requirement. "Data Protection by Design" invites controllers to adopt an appropriate technical and organizational design that includes pseudonymization in order to integrate data protection principles into their systems via design. "Data Protection by Default" mandates that by default, only personal data required for every particular purpose of processing are processed. For AI, this implies privacy-protection safeguards must be central to model design, data harvesting, and training processes, not something to be tacked on afterwards.

Security of Processing (Article 32)

Article 32 supports Article 25 in the sense that it expects the processor and controller to put in place "appropriate technical and organisational measures to ensure a level of security appropriate to the risk. These are, where appropriate, pseudonymisation and encryption of personal data, the giving to ensure the confidentiality, integrity, availability, and resilience of processing systems at all times, and a testing and evaluating process to determine the effectiveness of these on an ongoing basis. Instating such security controls is among the basic responsibilities of any company utilizing AI systems and an ongoing thread in cloud computing, when a large number of security services are on hand to address these requirements.

Data Subject Rights (DSRs) in the Age of AI

The GDPR also empowers the individual with a list of Data Subject Rights (DSRs) that among others include the right of access, correction, erasure ('right to be forgotten'), and data portability. Fulfilling these rights in the context of AI makes it serious technical challenges. For instance, calling for the right of erasure makes become radically challenging when the personal data of an individual is employed for machine learning model training. Removing raw data from a training set does not necessarily eliminate its effect on the parameters of a trained model. Methods for "machine unlearning" are a developing area of research but are not yet general or mature, and therefore there is a conflict between the right of erasure in law and the technical fact about the deployed AI system. Cloud providers provide data discovery and management features to assist in identifying and addressing data within DSRs, but it is the data controller who must comply with such requests, particularly with trained models.

The Roles of Controller and Processor (Article 28)

The GDPR establishes in particular the role of the 'data controller,' who decides the purposes and means of processing the personal data, and the 'data processor,' who processes the data on behalf of the controller. In the common cloud deployment of AI, the party that deploys or creates the AI system would be considered the data controller, and the Cloud Service Provider (CSP) would be the data processor. Article 28 mandates that arrangement must be underpinned by a contract (or Data Processing Addendum - DPA) that obliges the processor to provide adequate safeguards of putting in place adequate technical and organizational measures. Large CSPs such as AWS, Azure, and Google Cloud all offer foundation DPAs that state their undertakings to process data exclusively in accordance with the documented instructions of the controller and to help the controller fulfill its own GDPR responsibilities, including answering DSRs and keeping processing safe.

4.3 The EU AI Act: A Risk-Based Framework for Artificial Intelligence

While the GDPR governs the processing of personal data, the EU AI Act (Regulation (EU) 2024/1689) establishes a horizontal regulatory framework for the design, development, and deployment of AI systems themselves, regardless of whether they process personal data. Its central innovation is a risk-based approach, which calibrates the intensity of legal obligations to the level of risk an AI system poses to health, safety, and fundamental rights. [30]

The Tiered Risk Model

The AI Act classifies AI systems into four distinct risk categories:

a. Unacceptable Risk: The category consists of AI systems that clearly pose a danger to the safety, work, and human rights of people. These systems are strictly prohibited. Examples are public authority social scoring, real-time remote biometric identification in publicly accessible areas for law enforcement (with limited exceptions), and AI systems that use manipulative or exploitative approaches against vulnerable groups.

b. High-Risk: This is the most significant compliance category. It includes AI systems applied in some sensitive applications where they would have the potential to significantly affect individuals' lives. The Act aims at a line of such fields, e.g., AI as safety measures in critical infrastructure, in education and vocational training, in staff and workers management (e.g., CV-filtering software), in access to basic services such as credit scoring, in policing, and in administration of justice. Such systems are not prohibited but are subject to a complete regime of demanding conditions.

c. Limited Risk: These are computer systems with particular transparency risks. The most important obligation is to inform users that they are communicating with an AI system. This comprises chatbots, emotional detection systems, and systems generating "deepfakes." The output of these systems must be recognizable as artificially created or modified.

d. Minimal Risk: These are the overwhelming majority of AI systems, e.g., video games or AI-powered spam filters. These systems are not subject to statutory legal requirements under the Act, although they are induced to follow the voluntary application of codes of conduct.

The majority of the regulatory burden under the AI Act rests with deployers and suppliers of high-risk AI systems. They are under a series of lifecycle-prolonging obligations before they can put a system on the market and during its functioning. These conditions actually impose most sensible AI and MLOps principles in hard law:

- Risk Management Systems: Providers must maintain, implement, document, and sustain an ongoing and iterative risk management system. This process should identify, estimate, and examine the anticipated risks the AI system might pose to health, safety, or basic rights, and implement adequate risk management actions.
- Data and Data Governance (Article 10): The Act sets severe controls over data utilized to train, validate, and test high-risk AI systems. Data sets have to be

"relevant, representative, free of error and complete." That entails setting stringent data governance and management procedures, testing data sets for possible bias, and reducing such biases to avoid discriminatory decision-making.

- Technical Documentation & Record-Keeping (Articles 11 & 12): Providers are required to provide comprehensive technical documentation to prove compliance with the Act's requirements prior to putting a system on the market. Such documentation should be retained for the whole duration of the system. High-risk systems shall be designed in a way that they automatically generate records (logs) to provide an element of traceability of system functioning.
- Transparency and Provision of Information to Users: Operators of high-risk systems have to be given explicit user guidance, for example, information regarding the purpose, capability, limitations, and human control measures the system is required to implement.
- Human Supervision: High-risk systems need to be designed and developed so that they can be supervised efficiently by humans. This involves the imposition of controls by which a human supervisor would be able to comprehend the abilities and constraints of the system, be able to see its activity, and be able to intervene, override, or shut down the system when necessary.
- Accuracy, Robustness, and Cybersecurity: Systems need to be accurate, robust, and secure to a level that is commensurate with their purpose and that avoids harm throughout their lifecycle. This entails immunity from errors, faults, or inconsistencies, as well as protection from hostile attempts to modify the behavior of the system.

Phased Implementation Timeline

The AI Act is not operational at once. Its enforcement has a phased timing, as some provisions become effective earlier than the others. The prohibition on AI systems with unacceptably high risks started becoming effective in January 2025. General-purpose AI model providers will be held to requirements after that, and all the requirements for high-risk systems will start applying 36 months after entry into force of the Act, giving organizations a longer runway on which to prepare for such stringent requirements.

Both the GDPR and AI Act models of regulation, though differing in their core concern, embody a twofold compliance issue. GDPR is a matter of the legal and safe processing of personal information, and this requires security to safeguard that information. The AI Act, by contrast, targets the risk of the AI system itself, requesting it to exhibit quality, documentation, and resilience irrespective of the kind of data it is processing. This presents a possible gap of accountability: an AI system

can be technically GDPR-compliant processing anonymized data but in a high-risk environment could be classified as high-risk under the AI Act. This requires a dual-compliance approach covering both the lifecycle of data and the lifecycle of the AI model.

Finally, the two regulations have different business philosophies. GDPR's Data Subject Rights model is very reactive in nature, forcing a company to act in reaction to someone's request. In contrast, the AI Act's high-risk system model is very proactive. It demands a conformity assessment and the development of complex technical documentation prior to a system's approval on the market under the law. This moves the compliance burden from a reactive, administrative process to an active, design-and-engineering process, with legal and security requirements being intrinsic in the AI design process from the beginning.

4.4 The Cloud Computing Context: Security and the Shared Responsibility Model

Current AI technologies are designed and implemented for the most part not on-premises devices but in the expansive, stretchy terrains of public cloud infrastructures. Cloud-based reality by its very nature dictates the deployment of security, privacy, and compliance with regulations. The partnership between a cloud customer and CSP is regulated by an innovative framework called the Shared Responsibility Model. This is not just a technical standard; it is an operational and legal strategic framework, establishing security responsibilities, delegating responsibility, and imposing on the cloud customer to assume the primary responsibility for GDPR and AI Act compliance. [31] [32] [33]

Security of the Cloud vs. Security in the Cloud

The Shared Responsibility Model is the theoretical model of cloud security. It makes a universal assumption of a clear segmentation of work: the CSP should be responsible for cloud security, and the customer for cloud security.

- Security of the Cloud (CSP's Responsibility): This includes the physical infrastructure that supports all services provided by the cloud vendor. CSP is responsible for the security of the hardware, software, network, and physical facilities that form the cloud. This includes the physical data center security (access controls, monitoring), network resilience across the globe, hypervisor security (security of the virtualization layer), and security of the host underlying operating systems.
- Cloud Security (Customer's Responsibility): The customer's responsibility is on the basis of cloud services they deploy and applications they implement.

The customer is entirely responsible and under control with their data, including its classification and encryption. They also have the task of setting and managing platform-level security, i.e., the guest operating system (security patches), network controls (security groups, firewalls), application security, and most importantly, Identity and Access Management (IAM) for his or her users and services.

The allocation of responsibility varies based on the cloud service model utilized—Infrastructure as a Service (IaaS), Platform as a Service (PaaS), or Software as a Service (SaaS).

- IaaS (e.g., Amazon EC2, Azure Virtual Machines, Google Compute Engine): The customer is most responsible. The CSP supplies the underlying physical and virtualization layer, but all else is the customer's responsibility, such as the guest OS, middleware, runtime, data, and applications.
- PaaS (e.g., AWS Lambda, Azure App Service, Google App Engine): More is left to the CSP, who takes care of the underlying OS and runtime. More emphasis by the customer is on protecting their application code, data, and user access controls.
- SaaS (e.g., Microsoft 365, Google Workspace): The CSP owns most of the stack, e.g., the application itself. The customer is minimally involved in the administration of their data within the application and user access and permissioning configuration.

Whether using the model or not, the customer retains control of their data, its tagging, and the users who have access to it. This holds true for all the key CSPs, such as AWS, Azure, and Google Cloud.

Table 4.4.1: Responsibilities

Responsibility Area	**IaaS (e.g., Virtual Machine)**	**PaaS (e.g., Serverless Function)**	**SaaS (e.g., Hosted Email)**
Data & Content	Customer	Customer	Customer
User Access & Identity Management	Customer	Customer	Customer
Application Logic	Customer	Customer	Cloud Provider
Network & Firewall Configuration	Customer	Shared	Cloud Provider
Operating System	Customer	Cloud Provider	Cloud Provider

Virtualization Layer	Cloud Provider	Cloud Provider	Cloud Provider
Physical Servers & Storage	Cloud Provider	Cloud Provider	Cloud Provider
Physical Data Center Security	Cloud Provider	Cloud Provider	Cloud Provider

Table 4.4.1: Comparative Analysis of the Shared Responsibility Model across Service Types. This table synthesizes the division of responsibilities, illustrating how the customer's security burden decreases as they move from IaaS to SaaS, while responsibility for data and access remains constant.

4.5 Implications for AI Systems and Regulatory Compliance

The Shared Responsibility Model has far-reaching consequences for organizations that host AI systems and want to be in line with regulations such as GDPR and the AI Act. It is a risk-transfer mechanism, transferring the legal and operational risks of application-level compliance by contract from the CSP to the customer.

Under GDPR, the CSP is either a "data processor" processing solely the written instructions of the customer, or a "data controller". The CSP exists simply to offer a secure and compliant platform, as seen through their various certifications (e.g., ISO 27001, SOC 2) and contractually enforceable DPAs. But final responsibility for verifying personal data is processed securely and legally lies with the controller. If there has been a breach of data due to a customer making a cloud storage bucket publicly accessible through an oversight, the customer bears legal liability, not the CSP.

Similarly, under the AI Act, the organization that develops an AI system or places it on the market is defined as the "provider," while the entity using it is the "deployer". These are the actors who bear the legal obligations for high-risk systems. The CSP merely offers the base AI/ML services (e.g., Amazon SageMaker, Azure Machine Learning, Google Vertex AI) but is not the "provider" of the customer's tailored AI application based on the services. The customer is then held liable for performing risk assessments, data quality checks, generating technical documentation, and adding human oversight, among others.

This model can provide an "illusion of compliance." Companies might feel that if their CSP is compliant and certified against a number of standards, it will automatically be compliant for any application developed on top of this platform. This is a dangerous fallacy. The CSP compliance is security of the cloud; the customer needs to separately achieve compliance for their application in the cloud. This requires a strong level of cloud security and regulatory knowledge in the customer organization. They do not just outsource risk; they have to determine how to use the security and governance offerings offered by the CSP in order to satisfy their regulatory requirements.

4.6 Operationalizing the EU AI Act's Requirements for High-Risk Systems

The EU AI Act lowers principles of responsible AI at the high level to tangible, binding legal commitments for high-risk AI systems. When it comes to companies that utilize cloud platforms, compliance is a technical implementation issue and not a policy issue. This section gives a practitioner's analysis of how the fundamental

requirements of the AI Act can be met using the underlying services and tools of the three big cloud AI ecosystems: Amazon Web Services (AWS) SageMaker, Microsoft Azure Machine Learning, and Google Cloud Vertex AI. The data governance, technical documentation, human monitoring, and resilience requirements established by the Act are now no longer best practices but market access requirements, and the cloud platforms' set of tools gives what is required to meet these requirements.

Data Governance and Quality (Article 10)

Article 10 of the AI Act requires high-risk AI systems to be trained on, tested on, and validated on data sets that are "relevant, representative, free of errors and complete." Good data management and governance practices must also be employed to screen and reduce potential biases. This legislative requirement mirrors the MLOps principle of giving utmost importance to data quality as the foundation of any consistent model.

Amazon Web Services: Amazon SageMaker offers a combination of tools to meet these needs. SageMaker Data and AI Governance, on top of Amazon DataZone, enables organizations to build a centralized catalog of data and AI assets, improving discovery and collaboration.

SageMaker Data Wrangler provides a visual data preparation experience, which allows data scientists to visualize, clean, and transform datasets, so that they are "free of errors and complete." In addition, SageMaker Clarify provides the ability to inspect datasets for statistical bias, yielding metrics that enable organizations to detect and correct for prospective discriminatory effects before training, taking direct care of one of the major concerns of Article 10.

Microsoft Azure: Azure Machine Learning (Azure ML) provides end-to-end data governance features. Versioning and dataset tracking are supported for Azure ML data assets, with a transparent lineage being set from data to model. Integration with Azure Databricks and Unity Catalog provides a unified control plane for managing all data and AI assets in an organization with fine-grained controls and dynamic data lineage capture to the column level. The capability is invaluable for auditing intent so that training data was appropriate for the purpose intended by it, as required by the Act.

Google Cloud: Google Vertex AI Datasets is a centralized repository for data annotating and data management utilized in ML operations. In order to provide enterprise-scale governance, Vertex AI is configured with Dataplex, with data discovery features, quality tests, and lineage tracing across multiple data sources. Organizations can then define and implement data quality regulations and

guidelines in such a manner that only curated and validated data sets are utilized to train high-risk models, thus complying with the stringent requirements of the AI Act.

Technical Documentation and Traceability (Articles 11 & 12)

Articles 11 and 12 of the AI Act burden providers of high-risk systems with heavy documentation and record-keeping. Detailed technical documentation must be submitted by high-risk system providers prior to placing the system on the market, describing its intended purpose, functions, limitations, and compliance with the provisions of the Act. The systems must also be able to automatically record events to provide high traceability through the lifecycle of the systems.

Amazon Web Services: Amazon SageMaker Model Cards are specifically intended for that purpose. They create one place to record important model information, such as use cases, risk level, training, evaluation metrics, and fairness and bias comments. It directly corresponds to Article 11's requirements for technical documentation. In traceability, SageMaker Pipelines enables orchestration and logging of the complete ML pipeline from data prep to model deployment with an immutable audit trail being left behind. Combined with AWS CloudTrail, which records all API calls, it offers the full book-keeping required by Article 12.

Microsoft Azure: Azure Machine Learning Model Registry offers a repository to document, version, and track all the models. It keeps valuable metadata, such as the training job it was created from, its performance metrics, and its deployment status. This kind of registry is one of the vital building blocks in creating the technical documentation required. The MLOps functionality of the platform ingests the governance information throughout the life cycle, with lineage logging data tracking who deployed a model, why a change was made, and when deployed, in an excellent audit trail.

Google Cloud: Vertex AI Model Registry fulfills a similar function, offering a central registry for the management of the ML model life cycle. It enables versioning of the models and retaining associated metadata and test metrics.

Vertex AI Pipelines also automates and orchestrates ML pipelines and builds a larger execution graph that acts as an effective traceability tool. Every action in the pipeline, input and output, are traced in Vertex ML Metadata, establishing an auditable and fine-grained trace of the model development pipeline that can be utilized to confirm compliance.

Human Oversight

One of the basic premises of the AI Act is that high-risk systems must remain under effective human control. The law requires that systems be designed such that

they allow for effective human oversight, meaning the capability for a human to comprehend the state of the system and correct or halt its operation if necessary.

Amazon Web Services: Amazon Augmented AI (A2I) is a service designed specifically to put this principle into practice. It is a managed service to create human review workflows for ML predictions. A2I can be used to send low-confidence predictions or predictions in sensitive applications (e.g., loan applications) to human reviewers. This enables human judgment to be integrated directly into automated processes, giving organizations a clear mechanism for oversight and intervention.

Azure and Google Cloud: Both Google Vertex AI and Azure Machine Learning support the creation of human-in-the-loop (HITL) pipelines. Although they do not have a single, pre-packaged offering like A2I, their platforms consist of building blocks (i.e., data labeling tool, queuing service, and serverless function) that can be composed to create HITL pipelines. For instance, a model prediction is published to a message queue, which invokes a serverless function to show the result in a custom user interface for a human evaluator, whose ruling is recorded and fed back into the system.

Robustness, Accuracy and Cybersecurity

The AI Act requires high-risk systems to be highly accurate and resilient, and capable of resisting cyber attacks. Both requirements both fill in the completeness of the excellence of the AI model itself and the security of the operational environment it will be deployed into.

Accuracy and Robustness

Meeting the requirements for accuracy and robustness will require stringent testing and examination at every step of the model's life cycle.

Amazon Web Services: Amazon SageMaker Clarify offers capabilities to test models for performance and identify bias, ensuring they're fair and accurate.

SageMaker Model Monitor can be utilized to monitor the performance of deployed models continuously, identifying concept drift (when statistical characteristics of the target variable change) and data drift, the most important warnings for declining accuracy and stability.

Microsoft Azure: Responsible AI dashboard of Azure Machine Learning is one platform for model assessment. It contains error analysis capabilities, which flag clusters of data where the model is weak, and fairness metrics, which measure performance by demographic subgroup. These are the tools needed for widespread model checking under the Act.

Google Cloud: Vertex AI Model Evaluation provides the ability to calculate model performance against ground truth data across a broad range of classification and regression metrics like AUC, precision, and recall. For generative AI models, Gen AI evaluation service employs model-based metrics (where a strong "judge" model adjudicates quality) and computation-based metrics (like ROUGE and BLEU) in order to assess criteria like fluency, coherence, and relevance.

Cybersecurity

Securing an AI system against malicious attack is a floor requirement. This includes tapping into the wider set of cloud security services to establish a secure boundary around the AI workload.

AWS: AWS WAF (Web Application Firewall), AWS Shield for DDoS mitigation, and Amazon GuardDuty for sophisticated threat detection give protection depth to AI workloads facing the internet.

Azure: Azure Firewall, Azure DDoS Protection, and Microsoft Defender for Cloud each provide identical capabilities for network security, threat detection, and security posture management.

Google Cloud: Google Cloud Armor offers DDoS protection and edge-based WAF, while Cloud IDS offers network-based threat detection. These are supplemented by the Security Command Center, which offers a unified view of security and threats in the cloud ecosystem.

The broad and inclusive standards of the AI Act are an unstoppable driver towards uptake by mainstream mature Machine Learning Operations (MLOps) players. What used to be engineering "best practices"—data versioning, experiment logging, automated model surveillance, and complete documentation—are now legislatively mandated as legal standards for any entrant to the EU market with a high-risk AI system. The Act can therefore be seen not as stultifying innovation, but as a regulatory spur to reproducible, disciplined, and high-quality AI engineering. This cross-fertilisation between legal best practice and technical best practice also makes way for a new market in "Compliance-as-a-Service" solutions. The complexity of the Act, combined with the draconian financial sanctions for non-compliance, produces a high level of demand for mechanisms that automate and commoditize these obligations. Cloud providers are best positioned to address this need by packaging their current AI governance and security capabilities as end-to-end integrated compliance suites that offer a strong incentive for customers to construct their entire AI lifecycle, including their regulatory processes, in a single cloud environment.

Table 4.6.1: Requirements and solutions

AI Act Requirement	AWS Solution	Azure Solution	Google Cloud Solution
Data Governance & Quality	SageMaker Data Wrangler, SageMaker Clarify (Bias Detection), SageMaker Data & AI Governance	Azure ML Data Assets, Unity Catalog Integration, Responsible AI Dashboard (Data Analysis)	Vertex AI Datasets, Dataplex Integration, Vertex AI Model Evaluation
Technical Documentation	Amazon SageMaker Model Cards	Azure ML Model Registry, Responsible AI Scorecard	Vertex AI Model Registry, Explainable AI Reports
Traceability & Record-Keeping	SageMaker Pipelines, AWS CloudTrail	Azure ML Pipelines, Azure Monitor Logs	Vertex AI Pipelines, Vertex ML Metadata, Cloud Logging
Human Oversight	Amazon Augmented AI (A2I)	Custom Human-in-the-Loop (HITL) Workflows	Custom Human-in-the-Loop (HITL) Workflows
Accuracy & Robustness	SageMaker Clarify, SageMaker Model Monitor	Responsible AI Dashboard (Error Analysis, Fairness)	Vertex AI Model Evaluation, Continuous Evaluation
Cybersecurity	AWS WAF, AWS Shield, Amazon GuardDuty	Azure Firewall, Azure DDoS Protection, Microsoft Defender for Cloud	Google Cloud Armor, Cloud IDS, Security Command Center

Table 4.6.1: EU AI Act High-Risk System Requirements and Corresponding Cloud AI Platform Features. This table maps the principal obligations for high-risk AI systems under the AI Act to the specific services and features offered by the three major cloud providers to help organizations achieve compliance.

4.7 A Unified Framework for AI Security and Privacy

Compliance piecemeal is not feasible with the new regulatory environment. The GDPR and the EU AI Act, while differing in their underlying theme, complement each other. An effective security and privacy stance for AI hardware does require a unified framework to satisfy the requirements of both laws together. This can be achieved through effective use of the cloud hosting platform's root security controls natively. These technologies—data encryption, identity management, network

shielding, and logging—compose the technical foundation on which a dual-compliance approach can be constructed, uniting legal requirements into an interrelated collection of engineering fields.

The Interdependence of GDPR and the AI Act

Any separated model, where one group works on GDPR and another on the AI Act, is inefficient and is destined to fail. Both laws have a philosophical foundation of risk management and responsibility that finds major practical areas of commonality. For instance, the AI Act's stringent data governance controls in Article 10 requiring high-quality, representative, and unbiased data are directly supportive and confirming of GDPR's principles of data accuracy and fairness of processing. A company that has good data quality controls in place to meet the AI Act is also enhancing its GDPR standing.

Likewise, the "appropriate technical and organisational measures" of Article 32 of GDPR are a prerequisite of compliance with the AI Act's "cybersecurity" requirement in high-risk systems. A secure cloud environment shielded from unauthorized entry and data intrusions safeguards the personal data under GDPR and the integrity of the AI models under the AI Act. Thus, a strategic solution that treats these regulations as complementary pieces of a complete system of governance is the best way to comply.

Technical Safeguards for Dual Compliance

The foundation security services provided by the leading cloud vendors are the basis for this shared framework. They include the technical controls needed to meet both the shared requirements of GDPR and the AI Act.

Identity and Access Management (IAM)

The least privilege principle—associating the rights users and services need to get their job done with those rights—is fundamental to both data protection and security of AI systems. Cloud IAM services are the chief means of implementing the principle.

- AWS Identity and Access Management (IAM): Enables fine-grained policies to control access to all AWS resources, such as S3 buckets holding training data and SageMaker models.
- Microsoft Entra ID (previously Azure Active Directory): Offers an end-to-end identity and access management solution for Azure that supports RBAC for resources such as Azure Blob Storage and Azure Machine Learning workspaces.

- Google Cloud IAM: Offers a single place to manage permissions for all Google Cloud resources, enabling administrators to assign particular roles (e.g., "Vertex AI User," "Storage Object Viewer") to users, groups, and service accounts.

With IAM properly configured, a firm can guarantee that only authorized data scientists are allowed to access training data (a GDPR issue) and only authorized deployment pipelines can change production models (an AI Act issue).

Data Encryption Techniques

Encryption is particularly named in GDPR as a top-most technical measure for protecting personal data and is a cornerstone of cybersecurity in the AI Act. Cloud providers offer strong, multi-layered encryption capabilities.

- Encryption At Rest: All three leaders offer default encryption for data at rest within their core storage facilities, like Amazon S3, Azure Blob Storage, and Google Cloud Storage, and for block storage (EBS, Azure Disk Storage, Persistent Disk) and managed databases (RDS, Azure SQL, Cloud SQL). This encompasses encrypting data prior to disk writing, most often by way of AES-256.
- Encryption in Transit: Data being transmitted between cloud services or from the cloud to end-users is encrypted using transport-level encryption like TLS, by default.
- Customer-Managed Encryption Keys (CMEK): This advanced feature serves as an anchor in gaining complete control over data. Services like AWS Key Management Service (KMS), Azure Key Vault and Google Cloud KMS enable the customer to create, manage, and control their own cryptographic keys. The CSP utilizes these customer-managed keys to encrypt the data encryption keys securing the base data. This authorizes the customer to withdraw access to their data at and whenever they want, even from the cloud provider itself. This kind of functionality, in some settings also described as "crypto-shredding," is a strong instrument for asserting the right of erasure under GDPR and for safeguarding proprietary AI models as valuable intellectual property.

Network and Threat Protection

A secure and segregated network configuration is necessary to isolate AI systems from the outside world and block data exfiltration.

- Virtual Private Cloud (VPC): All three vendors provide VPC services (Amazon VPC, Azure Virtual Network, Google Cloud VPC) that enable customers to

allocate a logically separate segment of the cloud where they can deploy resources inside a virtual network that they define.

- WAFs and firewalls: AWS WAF, Google Cloud Armor, and Azure Firewall services enable network and application layer traffic filtering, defending AI endpoints against DDoS attacks, SQL injection, and web attacks. This is critical in ensuring the integrity and availability of high-risk AI systems.

Logging and Monitoring

Compliance can be evidenced through an irrevocable and complete audit trail. Both GDPR (for breach notification and accountability) and AI Act (for traceability as well as record-keeping) need robust logging.

AWS CloudTrail, Azure Monitor, and Google Cloud operations suite (previously Stackdriver) all offer detailed logs of every API call and admin activity on a cloud account. They monitor who did what, where, and when, and provide the blocks for any compliance audit or investigation to work off of. The governance of the future of AI will likely depend upon audit log inspection, and a properly configured logging and monitoring strategy will thus no longer be just a security best practice, but the central point for proving legal compliance.

Table 4.7.1: Articles and requirements

GDPR Article	Requirement Summary	Corresponding Cloud Controls/Services
Art. 25: Data Protection by Design & Default	Implement technical and organizational measures to embed data protection principles.	Encryption (Default & CMEK), Pseudonymization tools, IAM (Least Privilege), Data Loss Prevention (DLP) services.
Art. 32: Security of Processing	Ensure confidentiality, integrity, availability, and resilience of processing systems.	IAM, Encryption at Rest & in Transit, Network Security (VPC, Firewalls, WAFs), Threat Detection, Backup & Disaster Recovery.
Art. 33/34: Breach Notification	Detect and report personal data breaches to supervisory authorities and data subjects.	Logging & Monitoring (CloudTrail, Azure Monitor, Cloud Logging), Security Information and Event Management (SIEM) integrations.

Chapter III: Data Subject Rights	Fulfill requests for access, rectification, erasure, portability, etc.	Data Discovery & Classification tools (e.g., Amazon Macie), Database & Storage APIs for data access/deletion, IAM for access control.

Table 4.7.1: Mapping GDPR Articles to Cloud Security Controls. This table translates key GDPR articles into categories of cloud-native technical controls that help organizations fulfill their legal obligations.

Recommendations for Compliant AI Development and Deployment

According to the convergence of such legal frameworks and technical capacities, the following strategic suggestions must be adopted by organizations that design and implement AI systems on the cloud:

1. Adopt a "Compliance-by-Design" MLOps Life Cycle: Incorporate privacy and security considerations into each phase of the AI development lifecycle. Leverage cloud pipeline technologies to automate data validation, bias detection, model testing, and technical document generation in the CI/CD pipeline.
2. Develop a Cross-Functional AI Governance Framework: Form a governance committee with legal, compliance, security, data science, and business members. The committee will be tasked with categorizing AI systems based on the risk classification of the AI Act and monitoring the application of related compliance controls.
3. Prioritize Cloud-Native Security Services: Take full advantage of security services provided by the selected cloud vendor. As the default, implement a zero-trust model, require encryption everywhere (particularly CMEK for critical models and data), and implement fine-grained IAM policies on the least privilege principle.
4. Invest in Depth Logging and Auditing: Have logging services enabled across all applicable cloud resources and have logs stored for a duration long enough to satisfy regulatory and audit needs. Leverage cloud monitoring and security analytics features to regularly scan those logs for indications of threats or non-compliant behavior.
5. Encourage Ongoing Education: The technology and legal environments are rapidly changing. Companies must spend money on ongoing education for technical, compliance, and legal staff so they can keep pace with both the rules and the cloud solutions that exist to address them.

Chapter 5

Conclusions - Navigating the Digital Trilemma in an Age of Geopolitical Uncertainty

5.1 Introduction: The Unstable Equilibrium of Transatlantic Data Flows

This chapter is the capstone analysis of this thesis, distilling the overall argument that Europe's constant drive towards "digital sovereignty" is not a temporary policy choice but a structural and inherent response to an existential and recurring legal-political divide with the United States on the regulation of data. The research presented here has found that the 2025 transatlantic data environment is a mirage of calm and illusory political compromise covering profound, structural incompatibilities between two competing visions of law, rights, and state authority. Notwithstanding the prevailing balance the present harmony, then, is unstable by nature, as it is on a basis of political goodwill genuinely wearing away under the strain of increasing geopolitical tensions.

The analysis is framed against the backdrop of the significant geopolitical realignment which has occurred since the United States administration shifted in early 2025. This shift has not led to a new conflict but has instead intensified and revealed underlying differences which have long characterized the transatlantic data relationship. On one hand is the European Union, whose legal order, captured in the General Data Protection Regulation (GDPR), embeds an individual's right of protection of personal data as a constitutional right. On the other hand is the United States, whose legal order, specifically the Clarifying Lawful Overseas Use of Data (CLOUD) Act, is concerned with the accessibility of state and intelligence agencies to data for national security and law enforcement, wherever such data may be. All of this has established a mood of rudimentary and increasing uncertainty regarding the destiny of transatlantic digital commerce and collaboration.

This underlying tension finds voice in an inherent paradox. On 3 September 2025, the European Union General Court (GCEU) confirmed the legality of the EU-US Data Privacy Framework (DPF) in rejecting a legal complaint against its annulment. This decision has, at least temporarily, injected a sense of legal firmness and reliability

into the thousands of businesses that use this instrument for data transfers on a daily basis. But this judicial ratification is a long way from the unchecked legal phenomenon of the US CLOUD Act, which provides extraterritorial jurisdiction to the US authorities whose implementation collides head-on with the principles of the GDPR. Furthermore, the political ambiguity radiating from Washington dismantles the same executive guarantees on which the DPF agreement was founded and therefore its ultimate success is more than questionable.

This final chapter argues that this complex interaction of legal, technical, and political forces has put the European Union into a "digital trilemma". Europe is attempting simultaneously to meet three necessary, but seemingly contradictory, goals:

Digital Sovereignty: The ability to control its digital destiny, with information inside its borders governed by its own values and legislation, free of third countries' extraterritorial jurisdiction.

Technological Competitiveness: The imperative to access and leverage the latest and greatest cloud platforms globally to drive its economy, a market now dominated by a few US-based hyperscale providers.

Protection of Fundamental Rights: The ethical and legal requirement that the transfer of personal data to a third country be able to offer a level of protection to fundamental rights and freedoms which is "essentially equivalent" to that offered by EU law, consistently reaffirmed by the European Court of Justice (ECJ).

This trilemma will be the lens through which this chapter will analyze the position at present. It will critically explore the shaky legal basis for transatlantic data flows, review the technical and legal feasibility of the "sovereign cloud" offerings from American providers, and make an estimate of the strategic compulsions behind the establishment of a true European cloud environment. In conclusion, it will be posited here that engaging this trilemma is the usual European technology policy challenge in this period, and that the answer lies in a multi-level risk diversification ahead of time, long-term investment in an indigenous foundation, and uncompromising dedication to the supremacy of its own order of law. [28] [37]

5.2 The Enduring Legal Schism: Sovereignty, Surveillance, and the CLOUD Act's Long Shadow

The formation of transatlantic data transfer agreements is a history of juridical struggles and court invalidations. Both the Safe Harbor arrangement and its successor, the Privacy Shield, were struck down by the European Court of Justice in the landmark decisions of Schrems I and Schrems II, respectively. The substance of

both judgments was the failure of both schemes to provide EU citizens with some level of protection for their fundamental rights which is "essentially equivalent" to that which they enjoy within the Union, primarily due to the extent of US surveillance and impossibility of successful legal action for individuals. The EU-US Data Privacy Framework (DPF), approved in July 2023, was intended to overcome these specific flaws and represents the third attempt at establishing a stable legal basis for data flows. But appearances deceive, and closer examination shows that it is actually more politically negotiated truce than permanent legal resolution, with the foundations already being eroded by changing geopolitical alignments. [34] [35]

5.3 The DPF: A Political Solution to a Legal Problem

DPF's claim of adequacy is based on United States regulatory evolution, namely President Biden's Executive Order 14086 of October 7, 2022. In it, there were fresh, binding protections introduced geared toward restricting US intelligence agencies' access to information regarding individuals to the extent necessary and proportionate. Significantly, it also created a new, two-tiered EU data subjects' remedy system, subject to scrutiny by the Data Protection Review Court (DPRC), a court designed to be autonomous and able to issue binding directions against US intelligence agencies.

On September 3, 2025, the European Union's General Court of the European Union (GCEU) decided the case of Latombe v. Commission, rejecting French MP Philippe Latombe's application to annul the European Commission's adequacy decision for the DPF. Latombe had argued that DPRC was not independent enough and that US intelligence agencies remained free to conduct bulk data collection without adequate judicial oversight. The GCEU, in a decision that brought temporary respite to thousands of businesses, rejected these arguments on their merits. The Court held that the DPRC had adequate institutional and procedural protections for its independence and impartiality, mentioning that its judges hold domestic office for a fixed term, are excluded from holding a government office, can be removed only for cause, and whose findings are binding on the US government. The Court also spoke about bulk collection of data, including that the Schrems II decision did not require pre-judicial approval of all monitoring but required, at a bare minimum, post-judicial oversight, one requirement that is met in the new DPRC system.

In the wake of this judicial success, the stringency of the DPF is ipso facto eliminated. The structure of the entire edifice relies on the political promise of a US executive order, an instrument susceptible to change or revocation by a future administration. The GCEU judgment is a lagging indicator of already radically changed political reality. The role of the court was to determine adequacy of the DPF under existing legal and political realities as of the date it was adopted in July 2023. It

did not, and could not, look at subsequent weakening of political commitments that are the very foundation of the agreement. The US administration change in early 2025 has brought a new political dynamic, one marked by an expressed skepticism over the data protection and civil liberties protections erected by its predecessor. Its withdrawing members from the Privacy and Civil Liberties Oversight Board (PCLOB) constitutes a transparent policy change, one that undermines the good faith and trust implicit preconditions for any adequacy decision to qualify as robust and durable. Although the DPRC's structure is legally valid on paper, its functioning operational integrity and the surveillance environment in which it exists remain at the mercy of the whim of an executive branch less devoted to the principles codified by Executive Order 14086. This begins to produce a perilous dynamics of "compliance gap" for companies, where a framework judicially vetted is being politically unraveled systematically, rendering its eventual invalidation at the behest of a superior authority—the ECJ—a concrete and prospective possibility. [36]

5.4 The CLOUD Act's Irreconcilable Conflict with GDPR

Even assuming the DPF weathered the political tempest in its early days, it still does not touch on the underlying and structural legal clash: US CLOUD Act extraterritorial jurisdiction. Enacted in 2018, the CLOUD Act amended the Stored Communications Act (SCA) to specifically provide that US-headquartered tech companies must create data within their "possession, custody, or control" in order to comply with an issued US warrant or subpoena, irrespective of where that data is physically stored. For example, even when European company data is stored in Frankfurt or Paris data centers, because the service provider is a US-headquartered company like AWS, Microsoft, or Google, such data falls under the power of US law enforcement.

This is in express and irreconcilable conflict with the GDPR provisions. The GDPR, in particular, provides in Article 48 that any court judgment or administrative decision in a third country to require a controller or processor to disclose or transfer personal data is only recognizable or enforceable on the basis of an international agreement, for instance, a Mutual Legal Assistance Treaty (MLAT). The CLOUD Act was consciously drafted to circumvent the usually sluggish and cumbersome MLAT procedure, providing US authorities with a unilateral means to access information without inter-governmental agreement or judicial review within the EU.

This puts any European user of a US cloud provider in a position of continuing legal uncertainty, a paradigm "catch-22" scenario. If its cloud provider is presented with a CLOUD Act warrant, it is subject to US law and must comply. In doing so, it will be violating the GDPR for exporting data on an unlawful basis under EU law, putting its European customer at risk of significant fines and enforcement. In case it

does not comply, the provider itself will be legally penalized in the United States. The European Data Protection Board (EDPB) was adamant on this, stating that EU law subject service providers are not legally permitted to use CLOUD Act requests as the sole foundation for sending data to the US The outstanding tension de facto politicizes any assertion of "sovereignty" over data hosted with US providers, since final legal control lies not where data is stored, but where the parent company sits. [29] [37]

5.5 The Geopolitical Catalyst for "Schrems III"

The confluence of the DPF's fragile political underpinnings and the unresolved conflict of the CLOUD Act creates fertile ground for a new, successful legal challenge against the adequacy decision—a scenario widely referred to as "Schrems III." Privacy advocacy groups, most notably noyb, led by Max Schrems, have already signaled their intent to challenge the DPF, viewing it as a mere repackaging of the failed Privacy Shield with insufficient structural reforms to US surveillance law.

The political decisions of the newly elected US government in 2025 are a strong motive for such an effort. The February 2025 removal of three Democratic members of the PCLOB is a clear case in point. This action disempowers a significant oversight body responsible for ensuring privacy and civil liberties are protected within the US intelligence community. Likewise, withdrawn technology executive orders focused on other areas, such as risks related to AI, mirror a more general policy direction de-prioritizing just the kinds of protection upon which the European Commission was counting when it made the adequacy decision.

For the European Court of Justice, ultimately sitting in final judgment on any appeal, the most important question will be whether the overall US legal framework offers "essentially equivalent" protection. Political unraveling of the shields offered by Executive Order 14086, combined with the ongoing and unqualified application of the CLOUD Act, constitutes valid grounds for litigants that it does not. Max Schrems himself has publicly stated to be of the view that the DPF stands in imminent peril of repeal by the present US administration.

The effects of such an invalidation would be immediate and draconian. Transfers under the DPF data to the US overnight would become illegal, thrusting thousands of companies into a regulatory crisis reminiscent of the days following the Schrems II ruling. European businesses would be under huge pressure to review their entire IT infrastructure at pace, with a countdown to source alternative, GDPR-compliant solutions and potentially sever commercial relationships with key US service providers. For US tech companies, the effect would be equally seismic, with increased congressional risk in a critical European market and possible flight of

customers to sovereign European alternatives. This moment of current legal tranquility is thus best described not as a last stage, but as the eye of the inevitable legal storm to come.

5.6 The Hyperscaler's Gambit: A Critical Assessment of "Sovereign Cloud" Solutions

In response to the escalating regulatory pressure from the European Union and the persistent legal uncertainty surrounding transatlantic data flows, the dominant US-based hyperscale cloud providers—Amazon Web Services (AWS), Microsoft, and Google Cloud—have developed and heavily marketed a new category of offerings: the "sovereign cloud." These solutions are explicitly designed to address European concerns about data residency, operational control, and compliance with frameworks like the GDPR. They represent a sophisticated and resource-intensive effort to retain a commanding 70% share of the European cloud market in the face of growing calls for digital sovereignty. However, a critical assessment of these offerings reveals that while they introduce significant technical and operational safeguards, they ultimately fail to resolve the fundamental jurisdictional conflict that lies at the heart of the digital sovereignty debate.

5.7 Deconstructing the "Sovereign" Offerings

US hyperscalers' "sovereign cloud" offerings are not just repackaged data centers. They are architecturally separate environments designed to provide a high level of isolation from their international infrastructure. These are exemplified by products like the AWS European Sovereign Cloud, to be launched by the end of 2025 11; Microsoft Cloud for Sovereignty, offering tools and guardrails to regulated parties 12; and Google Cloud sovereign solutions, offered in collaboration with trusted European partners such as T-Systems in Germany and S3NS in France. All are supported by significant investment, with AWS putting €7.8 billion up to 2040 for its European Sovereign Cloud alone in Germany.

The key value proposition of these solutions is derived from a collection of advanced technical and operating mitigation designed to comply with rigorous European regulation:

Intrusive Data Residency: Underlying promise is absolute assurance that customer data is only hosted within the geographical confines of the European Union, and frequently within a designated member state. For example, the T-Systems Sovereign Cloud powered by Google Cloud ensures data is stored only in German data centers and under the jurisdiction of German data protection law. Likewise,

AWS's European Sovereign Cloud will be fully within the EU borders, with the first region being Brandenburg, Germany.

Operational Autonomy and Control: Among the most important innovations is the promise of operational autonomy. The providers guarantee that all day-to-day operations, access to the data center, technical support, and customer support will be in the hands of employees who live within the EU. AWS went the extra mile with a notice for a phase-wise transition such that its European Sovereign Cloud will be run by EU citizens, a move to safeguard legally and in reality the operations from outside EU interference.

Physical and Logical Isolation: These are not logically partitioned components of a global cloud. The AWS European Sovereign Cloud, for instance, is physically and logically isolated infrastructure that does not depend on non-EU infrastructure and does not have any mission-critical reliance on it. It is done to make sure there is resilience and avoid data or operational bleed-over.

Advanced Key and Encryption Management: Instead of simply accepting that data residency is not enough, the providers have good encryption controls. One key feature is that customers are able to keep control of their cryptographic keys. This is generally accomplished through external key management models, in which the keys are governed and kept by a trusted European partner outside of the US provider environment. Within the T-Systems and Google Cloud offering, T-Systems holds the encryption keys, and Google therefore cannot access the plaintext customer data AWS also accomplishes this through services such as AWS Key Management Service (KMS) with bring-your-own-key stores, which enable customers to host their own hardware security modules (HSMs). The AWS European Sovereign Cloud will even have a native dedicated, independent European trust service provider (EU-TSP) to handle the certificate authorities and key materials all within the EU.

These steps are not insignificant. They represent a major strategic change and a mass engineering endeavor to provide assurance and comply with the wording of numerous European statutes. The intention is to produce an environment in which, as a matter of operation, the provider does not have access to customer information. AWS's Nitro System, for example, contains technical restrictions that bar any operator, including authorized AWS staff, from reaching customer information on EC2 servers.

5.8 The Jurisdictional Achilles' Heel

In addition to these sweeping and technologically impressive protections, the whole structure of the US hyperscalers' "sovereign" products hangs on one determinant, but as yet undetermined, weakness: the jurisdiction of law of the parent

organisation of the US. The nub of it is that the jurisdiction of the US CLOUD Act is not premised on data's physical presence or on nationality of the operating entity undertaking a support role. Instead, its jurisdiction is derived through the corporate control doctrine; it is a term employed to refer to any information in a US company's "possession, custody, or control." Put simply, a German subsidiary of an American corporation, like AWS' GmbH in Germany to house its European Sovereign Cloud, is still under the legal jurisdiction of its American parent. If a US court grants a warrant under the CLOUD Act, the US parent company has a legal obligation to provide the data in question. That this data is located in Germany and is owned by EU citizens does not, according to US law, negate this obligation. While the provider has a right of appeal to a US court in respect of the request—a right provided for under the CLOUD Act itself—it is discretionary, sophisticated, and does not guarantee success. The ultimate decision lies with the US judiciary, not European authorities.

Such a fact has brought about a strategic decoupling of control of law from control of operation. The hyperscalers are making a strategic compromise to relinquish day-to-day operational control over their EU-based facilities, thereby satisfying most of the visible and tangible demands of European regulators and customers interested in matters like "Where is my data?" and "Who can touch it to support me?" But by maintaining ultimate legal and corporate authority in the United States, they guarantee that they will continue to be able to fulfill their US obligations, e.g., the CLOUD Act, if they are requested to do so. This generates an ingenious "compliance illusion," in which the underlying jurisdictional risk exists but is buried beneath layers of sound operational and technical mitigations. [37]

5.9 "Ringfenced" vs. "Sovereign"

In light of this ongoing jurisdictional risk, the phrase "sovereign cloud" for such products is a euphemism. Better descriptions would be "ringfenced," "compliance-optimized," or "regionally isolated" clouds. They are laboriously designed to be ringfenced on a technical and operational basis but not, and indeed not, on a legal and jurisdictional basis. Real digital sovereignty, in addition to physical and operational control, encompasses legal immunity from foreign countries' extraterritorial law.

It is of utmost importance to European organizations, especially in the public sector, critical infrastructure, and other heavily regulated sectors. Although these ringfenced solutions provide a higher level of security and compliance than non-ringfenced public cloud locations, they do not remove the underlying risk contained in the Schrems II judgment: the risk of non-EU governments' access to data in a way incompatible with EU fundamental rights. The decision to utilize these

services, then, remains one of risk tolerance, knowing that the overhanging, and possibly substantial, legal risk continues.

Below is a comparative summary of these services, outlining their characteristics and identifying the common jurisdictional issue.

Table 5.9.1: Providers and requirements

Provider/Offering	Data Residency Guarantee	Operational Control (Personnel)	Encryption/ Key Management Model	Stated CLOUD Act Mitigation Strategy	Remaining Jurisdictional Risk (Analysis)
AWS European Sovereign Cloud	Data and metadata stored exclusively within the EU.	Controlled by EU-resident, transitioning to EU-citizen, AWS employees located in the EU.	Customer-controlled encryption via AWS KMS; dedicated, autonomous EU Trust Service Provider for CAs.	Technical controls (e.g., Nitro System) to prevent operator access; legal challenges to requests.	High. The US parent company (Amazon.com, Inc.) retains legal control and is subject to the CLOUD Act, irrespective of the German GmbH's operational autonomy.
Microsoft Cloud for Sovereignty	Data residency within specified EU regions; tools to enforce residency policies.	Support and operations from EU-based personnel for specific workloads.	Azure Confidential Computing encrypts data in use; customer-managed keys and Azure Managed HSMs available.	Emphasis on robust compliance tools, transparency logs, and contractual commitments.	High. As a US-based corporation, Microsoft is subject to the CLOUD Act. The model focuses on providing customers with tools for compliance rather than offering structural immunity.
Google Cloud w/ T-Systems	Data stored exclusively in German data centers.	Technical support and operations provided by EU-based	External key management; T-Systems stores and manages encryption	Technical separation of duties; Google has no technical means to	High. While technically robust, the ultimate service provider is

		T-Systems staff.	keys outside of Google's infrastructure.	access plaintext data without keys held by T-Systems.	Google, a US company. Legal compulsion under the CLOUD Act could target the underlying infrastructure or the corporate entity, creating a legal conflict.

This analysis demonstrates that while the US hyperscalers have gone to great lengths to address European concerns, their solutions are fundamentally constrained by their national legal identity. They offer a powerful suite of tools to manage compliance risk, but they cannot offer the one thing that defines true sovereignty: freedom from foreign jurisdiction.

5.10 Forging a European Path: The Rise of Sovereign Alternatives and Strategic Imperatives

As the constitutional and geopolitical foundation for transatlantic data flows becomes increasingly precarious, an explicit alternative within Europe has made significant progress. It is more than a collection of alternative products; it is a new approach fundamentally to the cloud, one based on the principle of structural immunity and one compatible with a larger, multi-layered European strategy for achieving digital sovereignty. This approach involves not just the development of local cloud centers but also the establishment of universal norms, the unification of the continent's hardware base, as well as the creation of an active digital environment.

5.11 European Value Proposition: Structural Immunity

The central differentiating characteristic and essence of truly European cloud providers, like the French OVHcloud and Scaleway, is not any functional or service aspect but their legal position. They are constitutionally shielded from the extraterritorial application of foreign law like the US CLOUD Act because as companies with headquarters, seats, and activities bound by only the laws of the

member states of the European Union, they are under constitutional guarantees. They are legally bound primarily to the GDPR and the legal system of the EU.

This structural benefit provides a clear and definitive solution to the legal puzzle that surrounds consumers of US-based suppliers. In the case of a request for data from a third-country authority, e.g., a US-signed warrant, such European providers are not only capable but are obligated by law under the GDPR to resist it unless made through an accepted international process, i.e., a Mutual Legal Assistance Treaty (MLAT). This is what the GDPR Article 48 requires, and this is to invalidate third-country court orders that circumvent such accepted tools of legal assistance. This is not a policy decision or a market promise; it is a legal requirement hard-coded into their corporate DNA. To European players, especially the public sector and organizations dealing with sensitive individual or industrial information, this offers a degree of legal certainty and risk cover that "ringfenced" solutions outside the EU are not able to offer.

5.12 Beyond Infrastructure: Building a Sovereign Ecosystem

The advent of European cloud providers is not an exception. It is part of an overall, more comprehensive, and integrated European conception of its digital future. The vision understands that real sovereignty cannot be offered by the infrastructure alone, but has to include a complementary strategy across standards, hardware, and innovation.

Standardization and Interoperability (Gaia-X)

The Gaia-X initiative is one of the most important pillars of this strategy. It should be noted that Gaia-X is not, nor is it meant to be, a cloud provider per se, nor a try to create one "European hyperscaler." Gaia-X is a not-for-profit association responsible for creating the underlying architecture of a federated, interoperable, and sovereign European data space. Its focal deliverable is a "Trust Framework," an open set of standards, governance guidelines, and open-source software building blocks prescribing how data should be shared and services provided in an open, secure, and rights-protecting way.

The strategic objective of Gaia-X is to end the vendor lock-in cycle prevalent in today's cloud market, driven by vertically integrated, proprietary ecosystems. Through open standards and interoperability, Gaia-X seeks to establish a level playing field where European companies and citizens can transfer their data and applications securely between various compliant providers, and thus promote a competitive and pluralistic market of European services. As of June 2025, the Gaia-X project has come of age, having signed off on key versions of its Architecture and

Compliance documentation and formed a General Advisory Board to assist its global roll-out to nations like Japan, Korea, and Canada, and thus declare its aim to establish an international standard for trusted digital ecosystems.

Hardware and Technological Autonomy (EU Chips Act)

The European plan also recognizes the essential reality of the digital era: sovereignty is compromised when the hardware underneath is controlled by hostile supply chains with geopolitical influence. The pandemic and subsequent worldwide shortage of semiconductors, fueled by the COVID-19 pandemic and ongoing geopolitical tensions, served as a wake-up call. In response, the EU initiated the European Chips Act in 2022, a flagship industrial policy initiative supported by at least €43 billion of public investment, with the intention of supplementing this with private capital. The top-line target of the Act is ambitious: to double the EU's stake in the world value chain of the semiconductor market from approximately 9% in 2020 to 20% by 2030.

However, this goal is also confronted with harsh headwinds. In 2025, the European Court of Auditors reported that, at current levels of investment, the EU in 2030 will fall short of its goal and achieve only an 11.7% market share, several percentage points short of goal. This is a fundamental long-term weakness. The rapid pace of growth in AI and HPC is fuelling a bottomless demand for sophisticated semiconductors, i.e., GPUs. The European market for GPUs alone will expand from $10.6 billion in 2024 to $82.2 billion in 2034. Unless Europe is not reliant on non-democratic governments or geopolitical rivals for these strategic chips, industrial and digital sovereignty will continue to be in jeopardy, no matter the resilience of its cloud infrastructure or legislation.

Financing and Constructing Innovation (Digital Europe & Startup Initiatives)

The last component of the European strategy is the explicit promotion of a local digital ecosystem. This is also being achieved under both public and private initiative. The Digital Europe Programme, worth €8.1 billion over the 2021–2027 budgetary period, is a significant source of finance aimed at filling the gap between research being done in the university setting and market uptake, the long-term goal being to promote the extensive exploitation of digital technologies such as AI, data spaces, and cloud computing across the Union.

At the same time, European cloud providers are pushing hard to fill their platforms with the next generation of digital creators. Recognizing that the main strength of US hyperscalers lies not only in their infrastructure-as-a-service (IaaS) but

also in their huge, integrated platform (PaaS) and software (SaaS) services ecosystem, European providers are using startup programs as a strategic tool to develop a complementary ecosystem from scratch. OVHcloud's Startup Program offers substantial advantages, such as free cloud credits of up to €10,000 for scale-ups and up to €100,000 for early-stage startups, combined with technical assistance and guidance. Likewise, Scaleway's multi-level program offers up to €36,000 of credits, technical assistance, and access to a network of more than 800 startups.

These initiatives should not be viewed as publicity stunts. They are an important "seeding" strategy with the goal of building strong network effects that lead to vendor lock-in among US providers. By subsidizing and supporting emerging European technology firms on a sovereign infrastructure backbone from their earliest stages, these providers are making long-term investments to build a compounding effect. With increasingly innovative companies extending their solutions on European clouds, the ecosystem becomes richer and more appealing, attracting larger businesses in turn. This is a process that reinforces itself and one that eventually can offer a viable and legitimate alternative to US hyperscaler domination. The decision of the cloud provider is therefore raised from the level of a technical choice to that of strategic significance with far-reaching implications for risk, autonomy, and alignment with European policy objectives.

The following table outlines a strategic comparison, going beyond a bare list of features to emphasize the fundamental differences in value propositions within European and US providers.

Table 5.12.1: Requirements and providers

Strategic Criterion	European Providers (e.g., OVHcloud, Scaleway)	US Hyperscalers (e.g., AWS, Azure, GCP)
Jurisdictional Risk (CLOUD Act)	Structurally Immune: As EU-based legal entities, they are not subject to the CLOUD Act's direct jurisdiction and can legally oppose non-MLAT requests.	Mitigated but Present: As US-based corporations, they are subject to the CLOUD Act regardless of data location. "Sovereign" offerings mitigate operational risk but not the ultimate legal risk.
Data Sovereignty (GDPR Art. 48)	Full Compliance: Their legal structure ensures inherent compliance with the requirement to only honor third-country data requests made through formal international agreements.	Inherent Conflict: The CLOUD Act's mechanism for bypassing MLATs creates a direct and unresolved conflict with GDPR Article 48, placing customers in a state of legal ambiguity.

Ecosystem Maturity & Lock-in	Emerging and Federated: The ecosystem is less mature but is being actively built on open standards to prevent vendor lock-in, supported by initiatives like Gaia-X and startup programs.	Mature and Integrated: Possess vast, mature ecosystems of proprietary and third-party services, creating powerful network effects but also a high risk of vendor lock-in.
Commitment to Open Standards	Core to Strategy: Open-source standards (like OpenStack) are central to their value proposition, ensuring data reversibility and interoperability as a key differentiator.	Strategic Adoption: Support for open standards is often a feature, but the core business model relies on the integration of a proprietary service stack to drive customer retention.
Long-Term Strategic Alignment with EU Policy	Fully Aligned: Their entire business model is predicated on and aligned with the EU's strategic goals for digital sovereignty, data protection, and technological autonomy.	Commercially Aligned: Alignment is driven by commercial necessity to operate in the EU market. Their fundamental legal obligations remain with their home jurisdiction, creating a potential divergence from EU strategic interests.

5.13 Future Trajectories and Concluding Remarks: Towards a Resilient European Digital Future

The thesis developed during the course of this thesis and synthesised within the present concluding chapter delivers a definite and cogent set of conclusions about the state of digital sovereignty in Europe as of 2025. First, the rule of law for transatlantic data flows, apparently assured by the General Court's affirmation of the EU-US Data Privacy Framework, is an illusion and, in a harmfully contingent manner, dependent on a potentially volatile political environment. The fact that the framework is based on a US executive order, whose ground-level commitments are actually being eroded by a new government, makes it susceptible and a probable choice for a third consecutive invalidation at the hands of the European Court of Justice. The peace of today is not the solution but an interval.

Second, the "sovereign cloud" solutions built by US hyperscalers, as technologically advanced as they are, remain a type of strategic conformity. They are a gambit to meet the operational and residency requirements of European regulation but not to meet the underlying, intractable problem of foreign legal jurisdiction. By

differentiating operational control (the delegated responsibility of EU staff) from ultimate legal control (that which remains with the US parent company), they construct a strong but illusory story of sovereignty. They are most accurately described as "ringfenced" zones that avoid some risks but short of legal immunity under the US CLOUD Act to achieve complete sovereignty.

Hence, the thesis maintains that true digital sovereignty cannot be outsourced and cannot be secured by technical means whose subjects are subjected to competing and extraterritorial legal orders. It is a status that has to be institutionally ensured and based on a homogeneous and clear-cut legal order subject solely to the exclusive jurisdiction of the European Union and its member states.

5.14 The Strategic Imperative: Active Risk Diversification

Based on these observations, the strategic imperative for European public and private sector organisations is to move out of a passive, compliance-focused stance to one of positive policy of systemic risk diversification. To be dependent upon a single US hyperscale provider even with its most recent "sovereign" offer is to take on an unacceptable degree of legal, political, and commercial risk. The danger of an unexpected withdrawal of the DPF and the shadow over the horizon of the CLOUD Act pose a serious business continuity risk that cannot be avoided.

There needs to be a deliberate and thoughtful multi-cloud or hybrid-cloud approach as part of a cautious and robust strategy. It does not mean an immediate and complete turn away from US providers, whose well-developed markets provide undeniable benefit for non-sensitive workloads. Rather, it involves acknowledging the prevailing market reality and taking steps towards building resilience in advance. Organisations need to, as a strategy, segregate their data and loads based on their sensitivity and criticality. To the most sensitive private information, precious intellectual property assets, vital public services, and strategic industrial data, an escape to truly European sovereign providers would be a strategic imperative. This strategy voluntarily disseminates jurisdictional risk so that the organisation's most valuable digital resources are safeguarded by the strong and stable legal environment of the European Union.

5.15 Recommendations for a Multi-Pronged European Strategy

Reaching a viable digital future requires concerted and coordinated efforts on multiple fronts. The following is recommended by this thesis:

European and National Policymakers

Preserve Strategic Investments: preserve the robust financing and political backing of pillars such as Gaia-X and the EU Chips Act. Ensure that common standards for interoperability and a secure, regional supply of essential hardware are the long-term, sustainable pillar of digital sovereignty. The existing shortfall in achieving the objectives of the Chips Act must be addressed with haste, lest a fresh, yet even deeper layer of techno-dependency is established.

Implement Legal Requirements in International Bargaining: in future data transfer agreement negotiations, policymakers must insist on full legal reciprocity and preclude any arrangement that does not provide some standard of data protection that is verifiably and committedly "essentially equivalent" to the GDPR. The experience of the serial collapse of Safe Harbor and Privacy Shield must be fully learned: political assurances are insufficient without procedural legal reforms in the third country.

Support the Native Ecosystem: ongoing use of public budget tools such as the Digital Europe Programme to drive demand for sovereign cloud offerings and enable creating a European SaaS and PaaS ecosystem. Use public procurement strategies to set aside verifiably sovereign solutions for security-concerned government and public sector workloads.

Businesses and Public Sector Organizations

Perform Full Risk Assessments: perform a complete, jurisdiction-sensitive risk assessment of all cloud workloads and data assets. The evaluation should be superior to technical security but should involve an active consideration of the foreign provider's nationality, political, and legal risks. Data would need to be prioritized by sensitivity and labeled clearly as to what cannot be put at risk of foreign jurisdictional overreach.

Implement a Proactive Migration Plan: from the risk assessment, create a staged but purposeful migration plan for the most sensitive and business-critical data to providers who are solely under EU jurisdiction. This is not simply an exercise in compliance but a core component of modern risk management and business continuity planning.

Call for Openness and Interoperability: actively be part of and promote the new European ecosystem. In every cloud procurement, prefer and demand following open standards and data reversibility principles, not to be held hostage by vendors, and to enable the health and competitiveness of the envisioned Gaia-X federated digital marketplace.

5.16 Concluding Statement

The quest for digital sovereignty is no exercise of technological protectionism or withdrawal from international cooperation. It is a rational and natural reaction to a reorganized world. It is the pragmatic affirmation of Europe's right to maintain its own legal concepts, guard its economic sovereignty, and protect its democratic values within a world where the digital and the geopolitical are increasingly enmeshed. The way forward is long and full of hefty dangers, from stepping beyond primitive hardware dependencies to confronting the mighty network effects of established world behemoths. But the 2025 geopolitics risks have penned in more colorful language than ever why complacency and apathy are the biggest dangers of all. This thesis argues that by adopting a strategic, multi-level strategy that marries pragmatic short-term risk aversion with audacious, long-term investment in its own tech and rule environment, Europe can triumphantly resolve the digital trilemma and establish a genuinely resilient, competitive, and sovereign digital future.

Bibliography

[1] Curiousraj, "Balancing Innovation and Longevity: A Practical Guide to Software Stack Selection", Medium, 2025. [Online]. Available: https://medium.com/@curiousraj/balancing-innovation-and-longevity-a-practical-guide-to-software-stack-selection-10cdf41f4daa (cit. on p. 5).

[2] Apilover, "Best Tech Stacks Every Developer Should Know: A Comprehensive Guide", Dev.to, 2025. [Online]. Available: https://dev.to/apilover/best-tech-stacks-every-developer-should-know-a-comprehensive-guide-1pf5 (cit. on p. 5).

[3] Qovery, "AWS vs GCP vs Azure: Which Cloud Platform is Best for your Business", 2025. [Online]. Available: https://www.qovery.com/blog/aws-vs-gcp-vs-azure/ (cit. on p. 12).

[4] Veritis, "AWS vs Azure vs GCP: Cloud Cost Comparison for Enterprises", 2025. [Online]. Available: https://www.veritis.com/blog/aws-vs-azure-vs-gcp-cloud-cost-comparison/ (cit. on p. 12).

[7] AWS, "AWS Pricing Calculator", 2025. [Online]. Available: https://calculator.aws/ (cit. on p. 13).

[8] Azure, "Azure pricing calculator", 2025. [Online]. Available: https://azure.microsoft.com/en-us/pricing/calculator/ (cit. on p. 13).

[9] Google Cloud, "Google Cloud Pricing Calculator", 2025. [Online]. Available: https://cloud.google.com/products/calculator (cit. on p. 13).

[10] CloudZero, "Cloud Cost Management: The Ultimate Guide", 2025. [Online]. Available: https://www.cloudzero.com/blog/cloud-cost-management (cit. on p. 13).

[11] AWS, "Amazon EKS Pricing", 2025. [Online]. Available: https://aws.amazon.com/eks/pricing/ (cit. on p. 19).

[12] Azure, "Azure Kubernetes Service Pricing", 2025. [Online]. Available: https://azure.microsoft.com/en-us/pricing/details/kubernetes-service/ (cit. on p. 19).

[13] Google Cloud, "Google Kubernetes Engine Pricing", 2025. [Online]. Available: https://cloud.google.com/kubernetes-engine/pricing (cit. on p. 19).

[14] Microsoft, "Azure Container Apps pricing for consumption tier - Learn Microsoft", 2025. [Online]. Available:

https://learn.microsoft.com/en-us/answers/questions/2265079/azure-container-apps-pricing-for-consumption-tier (cit. on p. 19).

[15] Cast AI, "GKE Pricing Explained: How to Choose the Right Plan for You", 2025. [Online]. Available: https://cast.ai/blog/gke-pricing-explained-how-to-choose-the-right-plan-for-you/ (cit. on p. 19).

[16] Google for Startups, "Cloud", 2025. [Online]. Available: https://startup.google.com/cloud/ (cit. on p. 24).

[17] Google Cloud, "Startups", 2025. [Online]. Available: https://cloud.google.com/startup (cit. on p. 24).

[18] Microsoft, "Microsoft for startups Founders Hub benefits guide", 2025. [Online]. Available: https://partner.microsoft.com/en-us/partnership/founders-hub-benefits-guide (cit. on p. 24).

[19] AWS Startups, "Get AWS Activate Credits", 2025. [Online]. Available: https://aws.amazon.com/startups/credits (cit. on p. 24).

[20] Microsoft, "Introducing startup-friendly offers from trusted partners in Microsoft for Startups Founders Hub", 2025. [Online]. Available: https://www.microsoft.com/en-us/startups/blog/trusted-partner-benefits/ (cit. on p. 24).

[21] David Heinemeier Hansson, "We have left the cloud", Hey.com, 2023. [Online]. Available: https://world.hey.com/dhh/we-have-left-the-cloud-251760fb (cit. on p. 30).

[22] Sahid Jaffa (GEICO), John Hilt (GEICO), “GEICOs Year Long Journey to Realizing the Impact of an OCP Optimized Infrastructure”, 2024. [Online]. Available: https://www.youtube.com/watch?v=9ZbP-oTPrKI (cit. on p. 31)

[23] Efim Mirochnik, “How Ahrefs Saved US$400M in 3 Years by NOT Going to the Cloud”, 2023. [Online]. Available: https://tech.ahrefs.com/how-ahrefs-saved-us-400m-in-3-years-by-not-going-to-the-cloud-8939dd930af8 (cit. on p. 31)

[24] Akhil Gupta, “Scaling to exabytes and beyond”, 2016. [Online]. Available: https://dropbox.tech/infrastructure/magic-pocket-infrastructure (cit. on p. 32)

[25] Martin Fowler, "Microservices", martinfowler.com, 2014. [Online]. Available: https://martinfowler.com/articles/microservices.html (cit. on p. 33).

[26] K. Saquib, "The Evolution of Software Architecture: From Monolithic to Microservices", Medium, 2023. [Online]. Available:

https://medium.com/@ksaquib/the-evolution-of-software-architecture-from-monolithic-to-microservices-6a0ff15ce326 (cit. on p. 35).

[27] DZone, "From Monolith to Containers: Real-World Migration Blueprint", DZone, 2025. [Online]. Available: https://dzone.com/articles/monolith-containers-migration-blueprint (cit. on p. 35).

[28] Nicola Sfondrini, "Edge, Sovereignty And Sustainability: The New Cloud Trilemma For Global Enterprises", Forbes, 2025. [Online]. Available: https://www.forbes.com/councils/forbestechcouncil/2025/06/13/edge-sovereignty-and-sustainability-the-new-cloud-trilemma-for-global-enterprises/ (cit. on p. 5, 60).

[29] European Parliament and Council of the European Union, "Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)", Official Journal of the European Union, 2016. [Online]. Available: https://eur-lex.europa.eu/eli/reg/2016/679/oj (cit. on p. 42, 63).

[30] European Parliament and Council of the European Union, "Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending certain Union legislative acts (AI Act)", Official Journal of the European Union, 2024. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L_202401689 (cit. on p. 43).

[31] AWS, "Shared Responsibility Model", 2025. [Online]. Available: https://aws.amazon.com/compliance/shared-responsibility-model/ (cit. on p. 46).

[32] Microsoft, "Shared Responsibility in the Cloud", 2025. [Online]. Available: https://learn.microsoft.com/en-us/azure/security/fundamentals/shared-responsibility (cit. on p. 46).

[33] Google Cloud, "Shared responsibility and fate in the cloud", 2025. [Online]. Available: https://cloud.google.com/architecture/framework/security/shared-responsibility-shared-fate (cit. on p. 46).

[34] Court of Justice of the European Union, "Judgment in Case C-362/14 Maximillian Schrems v Data Protection Commissioner (Schrems I)", 2015. [Online]. Available: https://curia.europa.eu/juris/liste.jsf?num=C-362/14 (cit. on p. 61).

[35] Court of Justice of the European Union, "Judgment in Case C-311/18 Data Protection Commissioner v Facebook Ireland and Maximillian Schrems (Schrems II)", 2020. [Online]. Available: https://curia.europa.eu/juris/liste.jsf?num=C-311/18 (cit. on p. 61).

[36] General Court of the European Union, "Judgment in Case T-738/23 Latombe v Commission", 2025. [Online]. Available: https://curia.europa.eu/juris/liste.jsf?num=T-738/23 (cit. on p. 62).

[37] SIRIUS Project, European Union Agency for Criminal Justice Cooperation, "THE CLOUD ACT", 2022. [Online]. Available. https://www.eurojust.europa.eu/publication/cloud-act (cit. on p. 60, 63, 66)