`

Cloud Based Delivery Engineering with AI Service Definition

GCloud15

`

Introduction

Service Optimisation and Right-Sizing helps organisations improve the performance, resilience, and cost efficiency of live digital services as they scale toward operational maturity. The service applies evidence-based analysis across architecture, cloud infrastructure, delivery pipelines, and operating models to identify inefficiencies, bottlenecks, and excess capacity. Using monitoring, testing, and delivery metrics, it enables informed decisions on scaling, automation, and optimisation. The service supports the transition from reactive, manual operations to a controlled, factory-based model, providing predictable performance, improved service reliability, and transparent cost control aligned to real user demand.

The Challenges of Delivery Engineering with AI

Implementing Service Optimisation

Implementing Service Optimisation and Right-Sizing is inherently challenging because it requires changing how organisations understand, measure, and operate their digital services. The service moves teams away from assumption-based decision-making and reactive delivery toward evidence-led optimisation, which often exposes technical, operational, and organisational constraints that have previously been hidden.

A primary challenge is the **lack of reliable baseline data**. Many services enter optimisation with limited observability, fragmented monitoring, or inconsistent metrics. As highlighted in the decks, systems can appear “healthy” until failures accumulate under load. Without end-to-end telemetry, cost visibility, and performance data, early optimisation efforts must focus on establishing monitoring, alerting, and traceability before meaningful right-sizing decisions can be made. This foundational work is necessary but can initially slow perceived progress.

Another challenge is **legacy architecture and technical debt**. Services built in a “garage” model often rely on manual steps, hard-coded pipelines, tightly coupled components, and outdated dependencies. These characteristics make right-sizing risky, as changes in one area can have unintended consequences elsewhere. Optimisation therefore requires careful bottleneck identification, incremental refactoring, and validation through testing and controlled releases, rather than large-scale, one-off changes.

Cost optimisation presents its own complexities. Cloud spend is frequently distributed across multiple services, environments, and suppliers, making it difficult to attribute costs to outcomes. Right-sizing requires correlating usage, performance, and cost data to understand what capacity is genuinely required versus what exists due to historic over-provisioning or risk aversion. Introducing cost transparency can challenge existing assumptions and requires stakeholder alignment on acceptable trade-offs between cost, performance, and resilience.

A further challenge is **operational readiness and cultural change**. Moving toward a factory-based operating model requires teams to accept automation, standardisation, and governance embedded into delivery. This can initially feel restrictive to teams accustomed to bespoke practices or hero-led problem solving. Clear ownership, consistent ways of working, and agreed quality gates must be established to enable predictable optimisation without slowing delivery.

The **complexity of distributed ecosystems** also complicates optimisation. As shown in the decks, failures often originate outside the core system, such as in provider integrations or downstream services. Right-sizing must therefore consider the wider ecosystem, including third-party dependencies, message queues, identity services, and external APIs. This increases

`
the importance of cross-service observability, proactive alerting, and realistic performance testing.

In practice, Service Optimisation and Right-Sizing involves a structured set of activities to manage these challenges. Initial activities focus on establishing a factual baseline through architecture review, monitoring assessment, cost analysis, and delivery process evaluation. This is followed by workload and performance analysis using real usage data, testing evidence, and incident history to identify constraints and inefficiencies.

Subsequent activities centre on defining and prioritising optimisation opportunities. These include infrastructure resizing, scaling policy changes, pipeline improvements, automation of testing and deployment, and enhancements to observability. Recommendations are assessed against risk, cost impact, and implementation effort to avoid destabilising live services.

Implementation activities are deliberately incremental. Optimisations are introduced in small, controlled steps, validated through testing and monitoring, and adjusted based on observed outcomes. This reduces the risk of regression and builds confidence across delivery and operational teams.

Finally, a key challenge is ensuring **sustainability of optimisation**. Right-sizing is not a one-off exercise; demand, usage patterns, and service maturity evolve over time. The service therefore establishes ongoing metrics, governance, and review mechanisms so optimisation becomes a continuous capability rather than a reactive intervention.

Overall, the challenge of implementing Service Optimisation and Right-Sizing lies not only in technical change, but in aligning data, delivery, operations, and governance to support evidence-based decisions. When executed effectively, it enables predictable performance, controlled cost, and scalable live service operations, but it requires discipline, transparency, and sustained commitment across teams.

Key Benefits

The following benefits are expected from the service.

- Improves performance, resilience, and stability by aligning services to real demand and usage patterns.
- Reduces cloud and operational costs through evidence-based right-sizing and capacity optimisation.
- Provides clear visibility of service health, cost, and delivery risk across complex digital ecosystems.
- Enables predictable scaling and performance during peak demand and live service growth.
- Accelerates issue detection and resolution through proactive monitoring and observability.
- Supports transition from reactive delivery to a controlled, factory-based operating model.
- Improves confidence in live service readiness through measurable quality and performance signals.

- Reduces reliance on manual intervention and specialist knowledge through automation and standardisation.

`

- Enables data-led decision making across delivery, operations, and governance teams.
- Establishes a scalable, sustainable operating model for national or high-volume digital services.
- Improved confidence for transition to live services

Return on Investment

The following demonstrates the expected return on investment.

- Reduced cloud infrastructure spend through removal of excess capacity and inefficient resource utilisation.
- Lower operational support costs through proactive monitoring and faster issue detection.
- Avoided rework and remediation costs by identifying defects and performance risks earlier.
- Faster, safer releases reduce cost of delay and accelerate delivery of business value.
- Improved cost predictability enables better budgeting and financial control.
- Reduced dependency on specialist resources lowers long-term staffing and escalation costs.
- Optimised scaling prevents over-provisioning while maintaining peak-demand performance.
- Improved service reliability reduces incident-related business disruption and recovery effort.
- Evidence-based optimisation reduces risk of costly late-stage platform changes.
- Scalable operating model lowers marginal cost as service usage and demand increase.

Approach

The following describes the activities that are available with this service.

Service Optimisation and Right-Sizing begins with an assessment of the current service architecture, operating model, and delivery approach. This includes reviewing cloud environments, infrastructure configurations, deployment pipelines, monitoring coverage, and existing non-functional requirements to establish a factual baseline of performance, cost, resilience, and delivery maturity.

The service analyses live workload data, usage patterns, and demand profiles using monitoring, observability, and performance metrics. This identifies bottlenecks, over-provisioned resources, under-utilised components, and areas of technical debt that impact cost, reliability, or scalability. Cost and usage data is examined to understand spend drivers and opportunities for optimisation.

Performance, resilience, and operational behaviours are evaluated through testing evidence, incident data, and service telemetry. This includes assessing readiness for peak demand, failure scenarios, and live service operation. Gaps in monitoring, alerting, and observability are identified and prioritised.

`

Based on the evidence gathered, the service defines right-sizing and optimisation recommendations. These may include infrastructure resizing, scaling policy changes, pipeline and automation improvements, observability enhancements, and operating model adjustments aligned to a factory-based approach. Recommendations are prioritised by risk reduction, cost impact, and delivery effort.

Where agreed, the service supports implementation of optimisations, including configuration changes, automation updates, monitoring improvements, and validation of outcomes through repeatable testing and metrics. Changes are introduced incrementally to minimise risk and disruption.

The service concludes with clear reporting of outcomes, including cost savings, performance improvements, residual risks, and an optimisation roadmap. This enables informed decision-making, predictable scaling, and a sustainable operating model aligned to real service demand.

Training

Training is available and will take approximately 3 days.

Support

Support is available from 08:00 to 18:00 on working days in England. 1st

Line Support

1st line support will be available between the hours of 09:00 – 17:00 via the following channels;

- Web
- Phone
- email

2nd line support which will operate between the hours of 09:00 – 17:00.

Definition of SLA's (Will provide list)

- 0 – Critical – Within 1 Hour – Call back
- 1 – Severe – Within 2 Hours – Call back
- 2- High – 24 hours
- 3 – Medium – 48 hours
- 4 – Cosmetic – 72 hours