6. Scaling with Google Cloud Operations 透過 Google Cloud Operations 擴展規模

6.1 財務治理與管理雲端成本 (Financial Governance and Managing Cloud Costs)

6.1.1 雲財務治理的基本原則 (Fundamentals of cloud financial governance)

- **雲財務治理：** 組織用來管理雲支出的流程和控制措施，避免預算超支。
- **影響：** 人員、流程和技術。
- **人員 (People)：**
  _ 小型組織：一人可能負責所有方面。
  _ 大型組織：財務團隊負責財務規劃和顧問，技術和業務團隊提供資源使用建議。 * 需要跨財務、技術和業務功能的合作，或成立雲卓越中心 (Cloud Center of Excellence)。
- **流程 (Process)：**
  _ 每日/每週：監控和分析雲使用情況和成本。
  _ 每週/每月：財務團隊分析結果、成本返還和確定是否需要更改。 * 建立問責文化和跨團隊合作。
- **技術 (Technology)：** Google Cloud 提供內置工具來監控和管理成本，提高可見性、推動問責文化、控制成本和提供優化建議。

6.1.2 雲財務治理最佳實踐 (Cloud Financial Governance Best Practices)

- **確定雲成本管理者：** 最好由 IT 經理和財務控制員組成，建立全組織的成本問責文化，使用 Google Cloud 的財務治理政策和權限控制支出和成本查看權限，並使用預算和警報。
- **了解發票與成本管理工具的區別：** 發票是請求付款的文件，成本管理工具是追蹤、分析和優化雲支出的軟體。
- **使用 Google Cloud 的成本管理工具：**
  - 捕獲雲資源使用情況（使用者、目的和成本）。
  - 確定監控和管理成本的負責人，以及報告方式。
  - 定期審查報告（至少每週一次）。
  - 使用 Google Cloud 價格計算器 (cloud.google.com/products/calculator) 估算雲使用變化對成本的影響。

6.1.3 使用資源層次結構來控制訪問 (Using the Resource Hierarchy to Control Access)

- Google Cloud 資源層次結構包含四個級別：
    1. 組織節點 (Organization Node)：頂級，包含所有項目、文件夾和資源。
    2. 文件夾 (Folders)：可以包含項目或子文件夾。
    3. 項目 (Projects)：資源被組織到項目中。
    4. 資源 (Resources)：包括虛擬機、Cloud Storage 桶、BigQuery 中的表等。

- **策略 (Policy)：** 定義誰可以訪問資源以及他們可以對資源做什麼的規則，可在項目、文件夾和組織節點級別定義。

- **使用資源層次結構的優點：**
    - 細化的訪問控制：在不同級別分配角色和權限。
    - 權限繼承：簡化訪問管理。
    - 增強安全性和合規性：通過最小特權原則減少未經授權的訪問風險。
    - 強大的可見性和審計能力：追蹤訪問權限和變更。

6.1.4 控制雲資源消耗 (Controlling Cloud Consumption)

- **控制雲資源消耗的原因：** 節省成本、增加可見性和提高合規性。
- **Google Cloud 提供的工具：**
    - **資源配額策略 (Resource quota policies)：** 限制項目或用戶可以使用的資源數量。
    - **預算閾值規則 (Budget threshold rules)：** 當雲成本超過某個閾值時發出警報。
    - **雲計費報告 (Cloud Billing reports)：** 追蹤和了解已花費的費用，並提供優化成本的方法，可將計費數據導出到 BigQuery 進行詳細分析，並使用 Looker Studio 等工具可視化數據。
- **承諾使用折扣 (CUDs)：** 對於可預測的資源需求，承諾使用可獲得折扣價格。

總結：這份教材詳細說明了如何在 Google Cloud 中進行財務治理和控制雲資源消耗，包含人員、流程、技術、最佳實踐、資源層次結構和相關工具。

6.2 大規模營運卓越性和可靠性 (Operational Excellence and Reliability at Scale)

6.2.1 雲可靠性的基本原則 (Fundamentals of Cloud Reliability)

- **DevOps：** 強調開發和運營團隊之間的合作和溝通，以提高軟體交付的效率、速度和可靠性。
- **站點可靠性工程 (SRE)：** 結合軟體工程和運營，設計、構建和維護可擴展和可靠的基礎設施。
- **四個黃金信號：**
    - **延遲 (Latency)：** 衡量系統返回結果所需的時間，影響使用者體驗。
    - **流量 (Traffic)：** 衡量到達系統的請求數量，是當前系統需求的指標。

    - **飽和度 (Saturation)：** 衡量系統接近容量的程度，通常與性能下降相關。
    - **錯誤 (Errors)：** 衡量系統故障或其他問題的事件。
- **SRE 的三個主要概念：**
    - **服務級別指標 (SLIs)：** 顯示系統或服務性能的測量指標，例如響應時間、錯誤率或正常運行時間百分比。
    - **服務級別目標 (SLOs)：** 根據 SLIs 為系統性能設定的目標，例如系統在一個月內應有 99.9% 的可用時間。
    - **服務級別協議 (SLAs)：** 雲服務提供商與其客戶之間的協議，概述關於服務質量的承諾和保證，包括 SLOs、性能指標、運行時間保證以及違約的罰款或補救措施。

6.2.2 設計有彈性基礎設施和流程 (Designing Resilient Infrastructure and Processes)

- **高可用性 (High availability)：** 即使發生硬體或軟體故障，系統仍能保持運行和可訪問性。
- **災難恢復 (Disaster recovery)：** 在重大中斷或災難後將系統恢復到功能狀態的過程。
- **關鍵設計考慮因素：**
    - **冗餘 (Redundancy)：** 複製關鍵組件或資源以提供備份替代方案，提高系統可靠性。
    - **複製 (Replication)：** 創建多個數據或服務副本並分佈在不同的服務器或位置，確保容錯性。
    - **多區域部署 (Multi-region deployment)：** 跨多個地理區域分佈資源，提高彈性並減少服務中斷的風險。
    - **可擴展性 (Scalability)：** 構建可擴展的基礎設施以處理不同的工作負載和滿足增加的需求，使用自動擴展機制。
    - **備份和恢復 (Backup and recovery)：** 定期備份關鍵數據和配置，並存儲在地理上分離的位置，以防止區域停電或災難。
    - 定期測試和驗證這些過程，並實施監控、警報和事件響應機制。

6.2.3 使用 Google Cloud 現代化運營 (Modernizing Operations by Using Google Cloud)

- **可觀察性 (Observability)：** 從系統內的各種來源收集、分析和可視化數據，以了解其性能、健康狀況和行為。

- **Google Cloud Observability：** 一套綜合的監控、日誌和診斷工具，提供統一的平台來管理和了解應用程式和基礎設施的性能、可用性和健康狀況。

- **Google Cloud Observability 的組成：**

    - **Cloud Monitoring：** 提供雲基礎設施和應用程式的全面視圖，收集指標、日誌和跟踪，並提供警報策略。

- **Cloud Logging：** 收集和存儲所有應用程式和基礎設施日誌，提供實時洞察。
- **Cloud Trace：** 幫助識別應用程式中的性能瓶頸，收集延遲數據。
- **Cloud Profiler：** 識別應用程式使用的 CPU 能力、內存和其他資源。
- **Error Reporting：** 實時計算、分析和彙總運行中的雲服務崩潰情況。

6.2.4 Google Cloud 客戶服務 (Google Cloud Customer Care)

- **四個不同的服務級別：**
  - **基本支持 (Basic Support)：** 免費，提供文檔、社區支持、雲計費支持和 Active Assist 建議。
  - **標準支持 (Standard Support)：** 推薦用於開發中的工作負載，提供無限次訪問英語支持代表、故障排除、測試和探索，以及 Cloud Support API。
  - **增強支持 (Enhanced Support)：** 專為生產中的工作負載設計，提供 24/7 多種語言支持、更快的響應時間、技術支持升級和第三方技術支持。
  - **高級支持 (Premium Support)：** 專為具有關鍵工作負載的企業設計，提供最快的響應時間、客戶感知支持、專用的技術帳戶經理、Google Cloud Skills Boost 培訓平台學分、事件管理服務、運營健康檢查和客戶感知支持。
- 增強支持和高級支持計劃都提供可額外購買的增值服務。

6.2.5 支持案例的生命週期 (The Life of a Support Case)

1. 創建案例 (Case Creation): 客戶在 Google Cloud 控制台中創建案例，提供問題詳細信息並選擇優先級（P4 到 P1）。

優先級 (Priority)	**影響 (Impact)**	**初始響應時間 (Initial Response Time) (僅供參考，實際可能因情況而異)**
P1 (重大影響)	生產系統嚴重中斷，業務營運受到嚴重影響。	1 小時
P2 (高影響)	生產系統功能受損，但業務營運仍可勉強維持。	4 小時
P3 (中等影響)	非生產系統或部分功能受影響。	8 小時
P4 (低影響)	問題對業務營運影響不大。	24 小時

2. 分診 (Triage)： 團隊審查信息、了解問題並確定其嚴重性，並可能要求提供更多信息。
   - 客戶服務團隊審查客戶提供的資訊，以了解問題、確定其嚴重性以及對客戶業務營運的影響。

- 在這個階段，團隊可能會要求客戶提供更多資訊或澄清。

3. 分配 (Assignment)： 對於複雜問題，案例會被分配給具有相應專業知識的支持工程師。
    - 在許多情況下，客戶服務代表會直接解決問題。
    - 對於更複雜的問題，案例會被分配給具有相應專業知識的支持工程師。

4. 故障排除和調查 (Troubleshooting and Investigation)：
    - 團隊分析提供的資訊、審查系統日誌，並進行各種診斷測試，以找出問題的根本原因。
    - 根據問題的複雜性，此階段可能需要與其他內部團隊或專家合作。

5. 溝通和更新 (Communication and Updates)：團隊分析信息、審查日誌並進行診斷測試以識別根本原因，並可能與其他內部團隊或專家合作。
    - 團隊分析信息、審查日誌並進行診斷測試以識別根本原因，並可能與其他內部團隊或專家合作。
    - 團隊與客戶保持定期溝通，提供進展更新、分享發現，並在需要時要求提供更多信息或採取行動。

6. 升級 (Escalation)：升級用於標記流程中斷或案例陷入困境的情況，但並非總是最佳解決方案，對於高影響問題，確保案例設定為適當的優先級更重要。

7. 解決和緩解 (Resolution and Mitigation)：確定根本原因後，團隊致力於解決問題或提供緩解計劃，並可能與更高層次的支持或工程團隊協商。

8. 驗證 (Validation)： 實施解決方案或緩解計劃後，團隊與客戶合作驗證解決方案的有效性。

9. 關閉 (Closure)：客戶確認問題解決後，支持案例將被關閉，團隊提供解決方案總結、記錄採取的步驟，並提供預防措施或未來最佳實踐的建議。

10. 客戶會收到反饋調查。

總結：這份教材詳細說明了如何在 Google Cloud 中實現營運卓越性和可靠性，包含 DevOps、SRE、高可用性、災難恢復、可觀察性工具和客戶服務支持。