Treasury Board of Canada Secretariat

Secrétariat du Conseil du Trésor du Canada
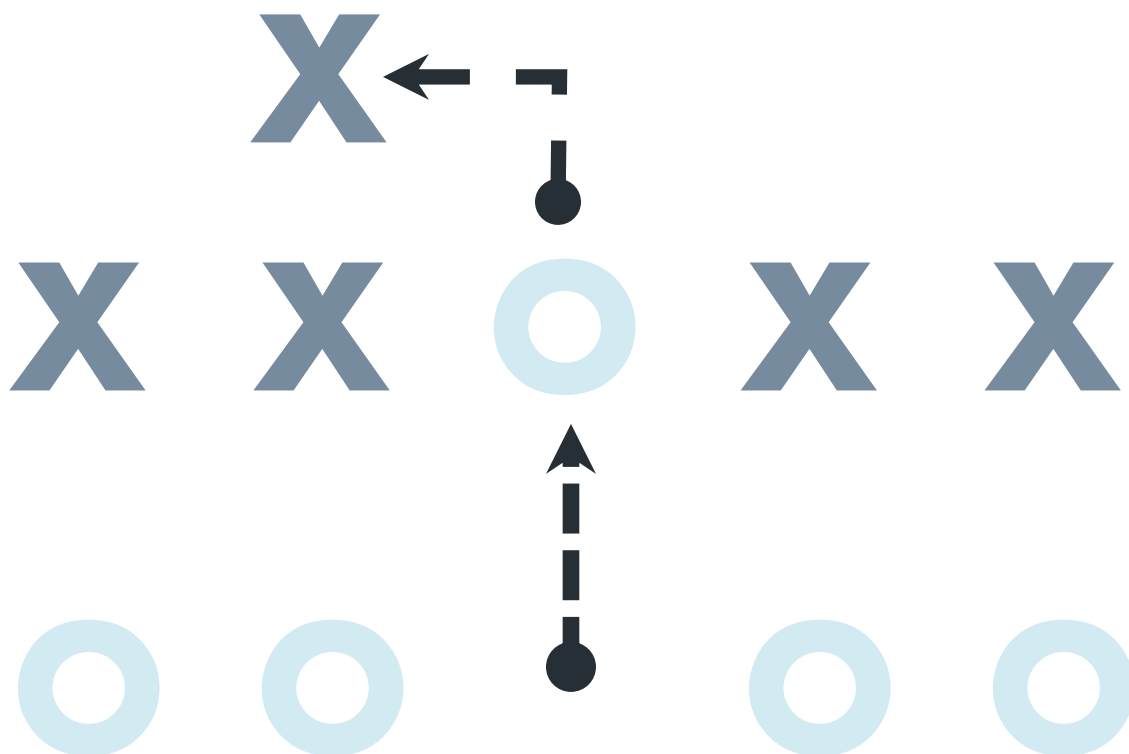
Canada

GC Cloud Cost Management Playbook

Published Version July 2022

Purpose

The purpose of this document is to:

- Support Government of Canada organizations adoption of cloud while applying financial due diligence while maintaining agility
- Provide a set of practices to maintain cost management in the cloud without negatively impacting agility and other cloud attributes
- Providing stakeholders with an overview of cloud cost management including financial & procurement officers

Context

Cloud computing provides on-demand, pay-as-you-go models for computing, storage, network, platforms, and business solutions. Cloud costs are fluid and directly related to the unit price of the resources multiplied by usage (cost = price x usage). This represents a fundamental shift from traditional, fixed, IT consumption models where the primary cost is associated with the purchase of hardware or licensing of software. In a cloud model, cost is a direct input into application architecture. Cloud spending is much more dynamic than traditional IT models as usage changes result in immediate cost changes. The feedback loop between usage and cost is immediate. Traditional spend controls rely heavily on process gates with people-intensive oversight. These traditional practices are a poor fit for a cloud environment as they can negatively impact the agility delivery teams desire when using cloud and lead to disrupted business value. Finance and procurement officers are not gatekeepers to cloud spend, but support delivery teams in establishing budgets for the delivery of their products. Delivery teams are directly responsible for cloud costs and must be prepared to track, analyze, and optimize the costs associated with their products and the usage of those products. These practices are often known as 'FinOps'. FinOps is a set of practices that encourage management & optimization of cloud spending without disrupting the agility and speed of cloud delivery.

FinOps

Often the term 'FinOps' is applied to cloud cost management. FinOps is a portmanteau of Financial Operations. FinOps is supported by a community lead [FinOps Foundation](). The foundation and technology consultancies have published a library of information on this topic. This guidance should not be read in isolation of that library of material. FinOps is an evolving area as cloud itself evolves.

GC Cloud Cost Management Playbook

Traditional Models Contrasted with Cloud

One can thinking about the shifting from traditional IT models to a cloud model using the analogy of shifting from digging and maintaining your own well vs connecting to the municipal water supply. Both deliver water, but each has a different model behind it. With a well, you must take on the responsibility of purchasing equipment, engineering, and building the well, and be responsible for maintaining the pumps and quality of the water. With a municipal water supply, you simply pay a monthly metered water consumption charge, and you are abstracted from the assets required to deliver the water. This, in many ways, is the shift that cloud has introduced. Traditional IT models, like digging wells, is a capital and asset intense delivery model. Cloud, like municipal water supply, turns the supply into a commoditized service.

The table below assumes that the pay-as-you-go (PAYG) model for cloud has been maintained. It should be noted that depending upon how your organization structures its agreement with providers, portions of this table may not apply. For example, if through a contract the consumer pays for years of services up-front, portions of this table will not apply. Furthermore, this table's scope are the cloud services themselves, and does not include capital costs for implementing the services or extending/building within these platforms.

| Traditional IT Models	Cloud Models
Traditional IT relies on a Capital Expense model (CapEX) where goods (software, hardware, facilities) are bought, maintained as an asset for a determined life span.	Cloud service are an Operational Expense, much like a utility such as electricity or municipal water, the degree of usage drives cost.
The degree to which an asset (hardware and software) is consumed has no material impact on cost. Once a purchase is made, the costs are sunk into the asset and there are no mechanisms to optimize those costs after purchase.	Consuming cloud resources results in costs. Turning off the usage of a service results in cost avoidance.
Traditional investments are typically periodic associated with the replacement of aging assets.	Cloud costs are in perpetuity until the service is turned off.
Traditional models have fixed costs.	Cloud costs are variable.
Traditional costs reporting is periodic, often annual as costs are largely fixed.	Cloud costs are real-time as they are variable
Traditional models have straight forward cost driver.	Cloud bills are complex often comprised of micro-transactions tracking each usage.
Purchasing of capital assets are centrally gated by procurement and financial organizations.	Cloud purchases are decentralized being made by delivery teams delivering digital products with the cloud.

Optimizing Cloud Costs

When contrasted against traditional IT models, cloud allows for more options for shaping spend. Levers in the form of optimization decisions can be taken to reduce waste and right-size consumption. Conversely, decisions can be made to spend more to get greater speed and scaling. Optimization choices can be plotted on a simple quadrant to illustrate the trade offs between time and cost as seen in figure 2. Optimization is focused on reducing waste and shaping spend. Continuous optimization may not lead to reductions in overall cloud spent over the long-term. Often, organizations continue to grow their cloud usage with time in the form of new services, new data, and new user experiences. The increased digitization of an organization's business will, over time, result in an overall increase in cloud spend. Optimization can help ensure that spend is well governed and not resulting in needless waste.

Feedback Loops

A key attribute of how costs are managed in the cloud depends upon the real-time feedback loop between using cloud and cloud costs. This simple feedback look drives financial behaviours that can lead to optimizing cloud usage to reduce needless spending.

Usage Changes:

- +/- user transactions
- +/- of cloud resources
- Architecture changes
- New features
- New user experiences

Product Delivery Teams
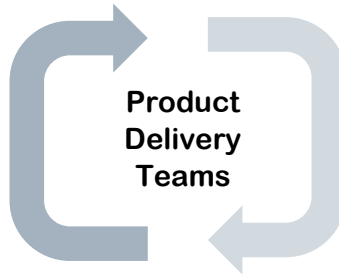
Cost Changes:

- +/- billing
- Cost forecast adjustments
- Optimization opportunities
- Provider pricing
- Volume discounts

Figure 1: Feedback loop between usage changes and cost changes drives optimization behaviours

Business Value, Not Lowest Cost

The goal of cost management is not to drive to the lowest cost but instead:

1. Optimize unnecessary waste
2. Delivery business value

Driving to the lowest cost possible can impact business value. Delivery teams and business owners may be willing to trade higher costs for quicker delivery. For example, a delivery team may decide to deploy an architecture that is not optimized because a simpler architecture may lead to faster delivery of a solution. The trade-offs between potential savings and the effort required to achieve those savings is illustrated below. Another example is prioritizing the launch of a new feature may be deemed higher that refactoring a piece of code that could lead to lower cloud costs. Therefore, opportunities for cost optimization become part of a product's feature backlog like. Opportunities to optimize a product for cost must be prioritized with other backlog items like new user experiences, new business rules, and other feature requests.

Managing cloud costs is not about driving to the lowest cost, but instead applying financial due diligence while driving business value for the organization.

Figure 2: Trade-offs between effort to optimize and the potential run savings gained for different cloud technology options

Plays

This document contains plays that organizations should undertake to exercise due diligence over cloud spend. This is not a comprehensive set of practices, but instead represent suggested, minimum practices.

01	Accountability
02	Visibility & Transparency
03	Cost attribution
04	Budgeting
05	Service Policies & Alerting
06	Optimization
07	Cost remediation
08	Governance
09	Set business metrics

GC Cloud Cost Management Playbook

⊞ 01 Accountability

Objective

Delivery teams[i] must be ready to assume cost management and optimization responsibilities. The implications of architectural designs and provisioning choices need to be understood by delivery team members.

Description of Activities

- ❑ Ensure delivery teams are aware of the impact of their consumption, configuration, and architectural choices have on costs
- ❑ Consider setting a cloud financial management (FinOps) training requirement for all delivery team members
- ❑ Ensure the basic concepts of FinOps practices and the guardrails listen in this document need to be understood by delivery team members
- ❑ Use the GC FinOps RACI (link required) as a template to tailor to your organization's roles
- ❑ Determine a centralized role within your organization who will manage cloud spend data and decisions with your organization's governance. This does not need to be a dedicated role, but part of a Cloud Centre of Expertise's responsibilities.
- ❑ Determine the role of financial officers in your organization's FinOps practices

References

- https://finops.world/en/organization-and-key-roles/
- https://finops.world/en/practices/card/1s1-assign-the-2-first-roles/

⊠ 02 Visibility & Transparency

Objective

Empower delivery teams with self-service access to cost reports. Delivery teams must have complete visibility into the services and usage that is driving cloud service costs associated with the delivery of their products. Be transparent with all stakeholders on cloud costs.

Description of Activities

❑ Provide reports and dashboards on past, current, and forecasted cloud spend to stakeholders

❑ Provide a standardized cost reports to all delivery teams, but allow delivery teams to tailor those reports to their needs

❑ Provide reports in such a way that a delivery team can isolate costs associated with the delivery of their products

❑ Maintain visibility over all cloud accounts and cloud services. If your organization uses multiple cloud providers, visibility over all providers and accounts will need to be maintained.

❑ Provide all stakeholders with transparency on cloud costs

❑ Determine how high in the organization cloud spend data should reach. For example, does the CIO or CFO have an interest in monthly reports?

References

• Cloud FinOps: Collaborative, Real-Time Cloud Financial Management by J.R. Storment and Mike Fuller

Automation

• Check to see that the billing dashboard or API is being invoked on a periodic basis

• Check that the developer and architect roles also have access to the billing console and APIs

03 Cost Attribution

Objective

Establish a structure for tagging cloud resources and attributing costs back to the lines of business and cost centres funding those costs.

Description of Activities

- ❑ Document and approve a standardized set of tags to identity how the cost of resources is attributed back to applications and lines of business for billing proposes
- ❑ Document and approved a resource hierarchy structure or groupings that support attribution of costs back to lines of business for billing purposes
- ❑ Use resource tagging and structures to act as reporting dimensions so that costs can be attributed back to lines of businesses and fund centres that fund the delivery of products
- ❑ Search and remediate any untagged resources or resources that exist outside of cost allocation structures
- ❑ Maintain the hygiene of resource:
  - ❑ Identify and remediate orphaned resources
  - ❑ Use automation through the implementation of service policies to enforce tagging hygiene where possible.
- ❑ For common cloud platform resources that cannot be attributed to a single product or line of business:
  - ❑ Determine a methodology for attributing common costs, for example:
    - ▪ Proportional; based upon relative percentage of direct costs (e.g., 10% of direct cloud spend)
    - ▪ Even split; split total amount evenly across products (e.g., common costs/number of products)
    - ▪ Fixed; like even split, but with a co-efficient to weigh the proportion allocated to each product. Total of all coefficients must add up to 1. (e.g., 0.3 of common costs)
  - ❑ Determine a show-back/charge-back methodology for making common costs visible to each product

References

- https://aws.amazon.com/blogs/aws-cloud-financial-management/cost-allocation-basics-that-you-need-to-know/
- Cloud FinOps: Collaborative, Real-Time Cloud Financial Management by J.R. Storment and Mike Fuller

Automation

- Check for resources that do not have approved cost allocation tags
- Use service policies to enforce mandatory cost allocation tags

04 Budgeting

Objectives

Delivery teams establish product budgets in collaboration with financial officers and undertake activities to track actual costs against forecasts.

Description of Activities

- ❑ Delivery team will work with financial officers to establish a budget for their product based upon estimated cloud hosting costs
- ❑ Budgets should be broken down by billing period. Alerting should be implemented to notify delivery teams when a pre-determined percentage of the monthly forecasted budget is spent,
- ❑ Delivery teams should report actuals against forecasts to determine if the 'burn rate' is within forecast.
- ❑ Governance should be documented to include process and criteria for increasing budgets.
- ❑ Governance of budget should allow for unplanned usage and costs, but also take corrective action to help delivery teams with repetitive unplanned spend.

References

- Cloud FinOps: Collaborative, Real-Time Cloud Financial Management by J.R. Storment and Mike Fuller
- https://aws.amazon.com/blogs/aws-cloud-financial-management/beginners-guide-to-aws-cost-management/
- https://docs.aws.amazon.com/cost-management/latest/userguide/budgets-create.html

Automation

- Check that the cloud service providers' features for setting and alerting on budgets have been configured

GC Cloud Cost Management Playbook

▣ 05 Policy & Alerting

Objectives

Use alerts to provide early warning when key financial milestones are met, such as percentage spend of total budget. Use service policies to enforce key financial guardrails such as adherence to tagging standards.

Description

❑ Use cloud platforms' ability to alert users of financial events as a method of automating governance activities

❑ Use service policies to enforce adherence to key financial guardrails such as tagging and automating turning off services, in particular for test and development environments. Financial guardrails differ from the GC Cloud Guardrails as they are department specific and focused on financial governance.

References

• Cloud FinOps: Collaborative, Real-Time Cloud Financial Management by J.R. Storment and Mike Fuller
• https://cloud.google.com/billing/docs/how-to/budgets
• https://docs.microsoft.com/en-us/azure/cost-management-billing/costs/cost-mgt-alerts-monitor-usage-spending
• https://docs.aws.amazon.com/AmazonCloudWatch/latest/monitoring/monitor_estimated_charges_with_cloudwatch.html

Automation

• Set alerts for accounts or resource groups based to notify delivery team members when 25%, 50%, and 75% (or whatever values best suit your context) of the team's total budget has been consumed
• Set alerts for excessive percent changes in costs occurring in a short period of time

06 Optimization

Objectives

Delivery teams should continuously analyze opportunities for optimizing service usage and architectures for cost efficiencies. Centralized FinOps teams can discover optimization opportunities to help delivery teams make optimization decisions.

Description of Activities

- ❑ Delivery teams will review cost and billing reports for possible optimization opportunities and waste
- ❑ Determine the benefit verses cost of implementing optimization strategies
- ❑ Opportunities to optimize architectures and usage should be added to the product's feature backlog to be prioritized along-side user feature requests and technical debt remediation. The goal is not the cheapest solution at all cost. A higher cost may be tolerated if greater deliver speed can be achieved.
- ❑ Identify possible sources of waste such as resources and resources with low usage rates
- ❑ Right-size resources, such a virtual-machines, by analyzing performance metrics. Move to smaller instances when possible.
- ❑ Use cloud providers' cost optimization services to review recommendations and decide whether to act upon them
- ❑ Rely on horizontal scaling to increase resources during peak periods.
- ❑ Align capacity and demand using elasticity of horizontal and vertical scaling
- ❑ Turn off resources not used outside of core hours such as development and testing environments.
- ❑ Analyze usage patterns to identify opportunity to migrate to reserved instances to achieve cost efficiencies
- ❑ Use service with greater ability to reduce pay-per-use costs. For example, moving from virtual machines to serverless platforms will provide greater cost granularity, but also consider the cost of refactoring may nullify those cost efficiencies gained.

References

- Cloud FinOps: Collaborative, Real-Time Cloud Financial Management by J.R. Storment and Mike Fuller
- https://aws.amazon.com/aws-cost-management/aws-cost-optimization/
- Maximize Business Value with Cloud FinOps, Google Cloud, https://services.google.com/fh/files/blogs/cloud_finops_paper.pdf

Automation

- Report on the number of features requests in product backlogs are related to cost optimization.
- Automate the periodic decommissioning of resources for development and testing environments outside of hours they are being used, for example, core hours.
- Report on the number of compute instance and storage provisioned and trends over time.

GC Cloud Cost Management Playbook

⊞ 07 Remediation

Objectives

When cloud costs grow beyond forecasts, implement an action plan to analyze current usage and reduce waste.

Description of Activities

When unplanned spend grows, implement a plan to remediate spending growth:

- ❑ Report on unexpected cost growth to governance
- ❑ Present an action plan to governance for addressing unexpected cost growth
- ❑ Obtain authority from governance to execute on an action plan to remediate unexpected cost growth
- ❑ Bring together implicated stakeholders to form a tiger team
- ❑ Hold blameless retrospectives as a method of learning from accidental cost overruns with the intent to improve future practices rather than focus on assigning blame
- ❑ Perform basic analysis. Determine the source of unplanned spend increases
- ❑ Trace back spend growth to the source business activities
- ❑ Set achievable targets for targets for spend reduction. Look for opportunities to reduce waste and optimize usage
- ❑ Implement spend reduction activities

References

- Cloud Cost Optimization, Gartner, Mike Brown
- Maximize Business Value with Cloud FinOps, Google Cloud, https://services.google.com/fh/files/blogs/cloud_finops_paper.pdf
- https://newrelic.com/blog/best-practices/blameless-retrospectives

GC Cloud Cost Management Playbook

08 Governance

Objectives
Design and implement a governance to review and act upon cloud cost analysis

Description of Activities
- ❑ Encourage cloud cost transparency through reporting, including:
  - o Costs by provider and by product
  - o Shared costs
  - o Spend against budgets and forecasts
  - o Trends over time
  - o Top and low performing delivery teams/products in terms of cost optimization
- ❑ Create processes for monitoring and taking corrective action, but without impacting the speed of delivery and agility of cloud
- ❑ Decide on financial guardrails and alerting all delivery teams and products must adhere to
- ❑ Cost governance does not require a dedicated governance vertical, but may be integrated into other governance activities such as performance, architecture, and security
- ❑ Document the roles and responsibilities associated with cloud cost management
- ❑ Favor guardrails over gates for enforcing policy

References
- Cloud FinOps: Collaborative, Real-Time Cloud Financial Management by J.R. Storment and Mike Fuller
- https://aws.amazon.com/blogs/enterprise-strategy/introducing-finops-excuse-me-devsecfinbizops/

09 Business Metrics

Objectives

Increased digitization of services and automation will drive up cloud costs over time that no amount of optimization efforts will be able to reverse. Business owners need to be explained cloud costs in terms of the business being delivered using cloud services. Delivery teams must be able to report on cloud costs in the context of business metrics. The unit economics, i.e., the cost per unit, must be calculated and reported.

Description of Activities

❑ Collaborate with business owners to understand their digital service deliver metrics (e.g., number of inquiries, wait times, number of transactions, speed of transactions, timeliness of transactions, total value of transactions, cost effectiveness, etc...)

❑ Report on cloud costs in terms of digital service delivery metrics (e.g., the average cost of serving an inquiry is X dollars)

❑ Express technology investments in terms of service deliver metrics improvements (e.g., increased investment will result in increased ability to serve more inquiries)

❑ Report on cost and digital service delivery metrics trends over time

References

- Cloud FinOps: Collaborative, Real-Time Cloud Financial Management by J.R. Storment and Mike Fuller
- https://aws.amazon.com/blogs/aws-cloud-financial-management/unit-metrics-in-practice-lessons-learned/

GC Cloud Cost Management Playbook

Use Cases

Treasury Board Secretariat – Optimization with Reserved Instances

Treasury Board Secretariat migrated its entire portfolio of 120 plus applications to the cloud in 2020. In that time they have gain considerable experience with optimizing their cloud usage. Moving to Reserved Instances was one way TBS optimized its use of Virtual Machines. By lock-in to a three-year purchase of VMs, a 60%, or greater, savings was achieved over pay-as-you-go models (PAYG). TBS has calculated, that in most circumstances, the break even point of a reserved instance is 15 months. This means that even if the VM is only used for 15 months of the 36 total purchase period, cost savings occur over a PAYG model. For TBS, who automated the scheduled shut-down and start-up of development and test environments to save money, moving to RIs has achieved greater cost savings than scheduling and with lower complexity.
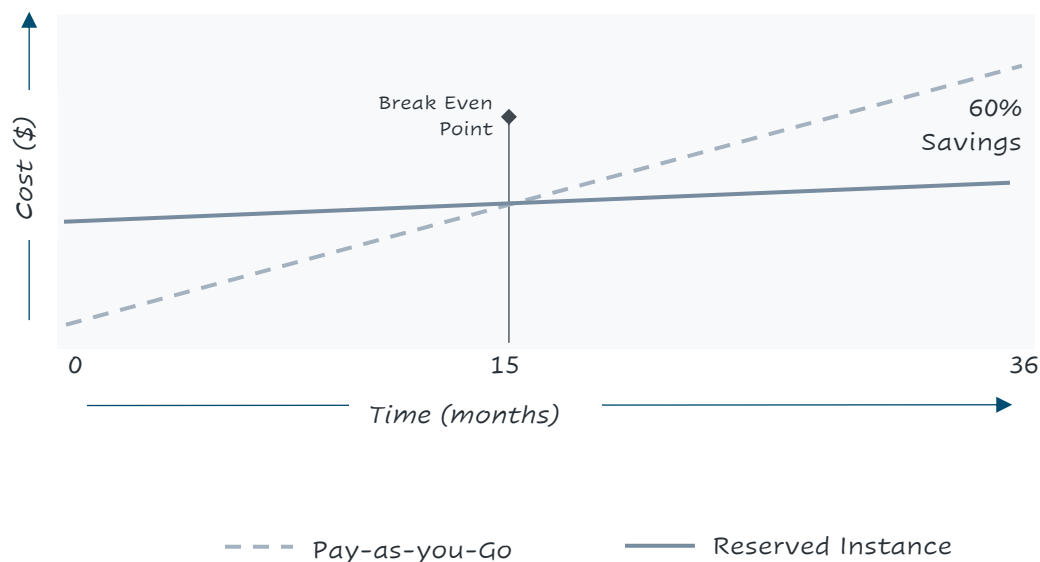
Figure 3: Point where pay-as-you-go model and reserved instance model break even

Next to migrating to reserved instances, TBS has also found setting and executing log retention periods are a considerable source of savings. Setting a retention policy for logs and then periodically cleaning up those logs that have exceeded the retention policy from block storage and PaaS yields cost avoidance.
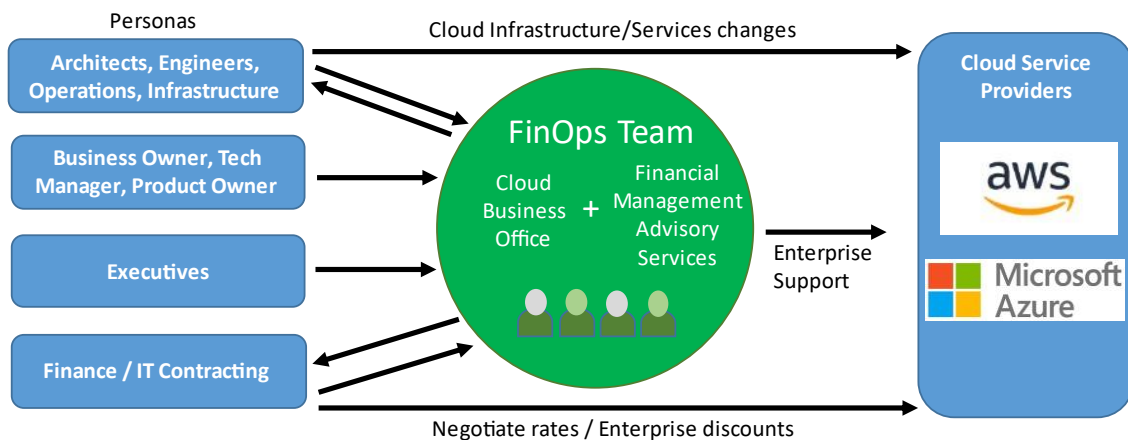
Canada Revenue Agency – Budget, Allocation, Payment

The Canada Revenue Agency (CRA) has assembled a cross-functional FinOps using stakeholders from their Cloud Business Office (CBO) and Financial Advisory Services. This team undertakes the work of forecasting, budgeting, and tracking actual expenses.

The process of forecasting, allocation, and payment at CRA is described as follows:

1. Annually, the Cloud Business Office (CBO) forecasts the spend for Azure and AWS, with a breakdown per Cloud Business Requirement. Adjustments are made during the year based on need, avoiding Q4 when possible
2. CBO makes a request to Central IT for commitment (funds added to requisition/order in CRA's SAP system)
3. IT Contracting puts in place the Cloud Brokering Service Orders per Cloud Business Requirement selecting the Pay-as-you-go model (i.e. you get invoiced for what you use, no upfront costs or purchase commitments)
4. Monthly invoices are reconciled and paid from Central IT's budget
5. Finance charges back actual costs to business owners as per cost allocation tagging, plus the 10% Cloud Brokering Service. An internal transfer of funds is recovered to Central IT
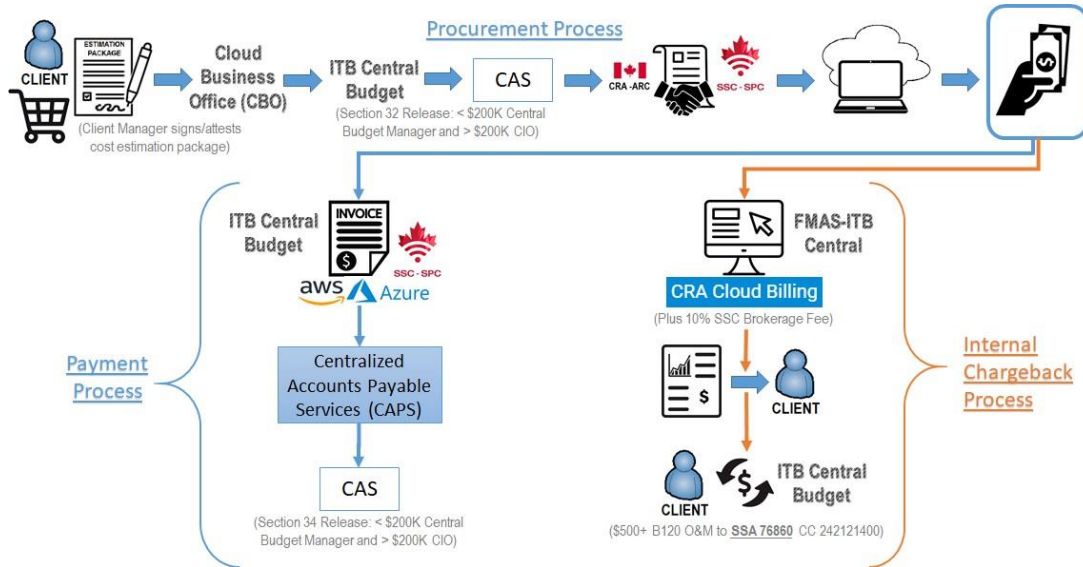
Notes:

1. Chargebacks are quarterly starting at P3 and then monthly in Q4
2. FY-end funding transfer cut-off constraint: midway through March, finance forecasts the anticipated spend for charging back, risk managing the delta
3. Shared Costs chargeback model is being developed in 2022

Statistics Canada – Finops Organization

When Statistics Canada committed to migrating its application portfolio to cloud, they underwent a large change management not only in technology, but also in governance and practices. Part of that was investing in FinOps and establishing a FinOps organization. Statcan brings together people from the financial and technology organizations to monitor, management, and optimize cloud spend.
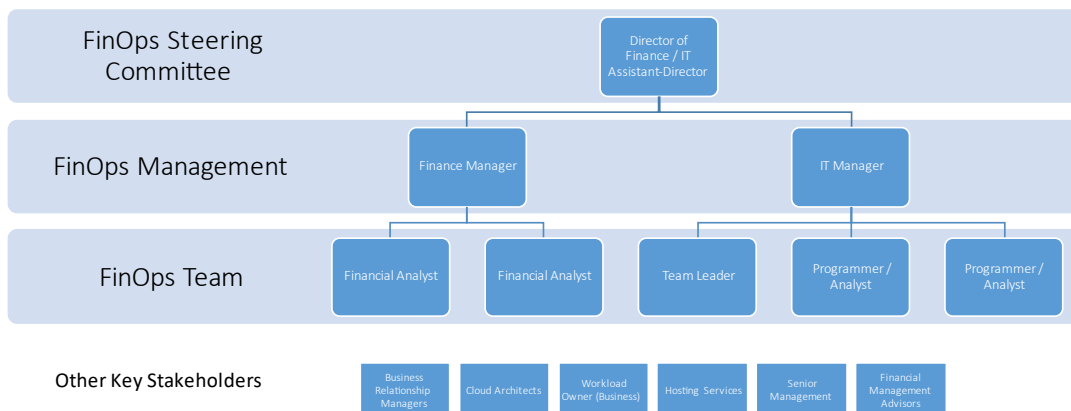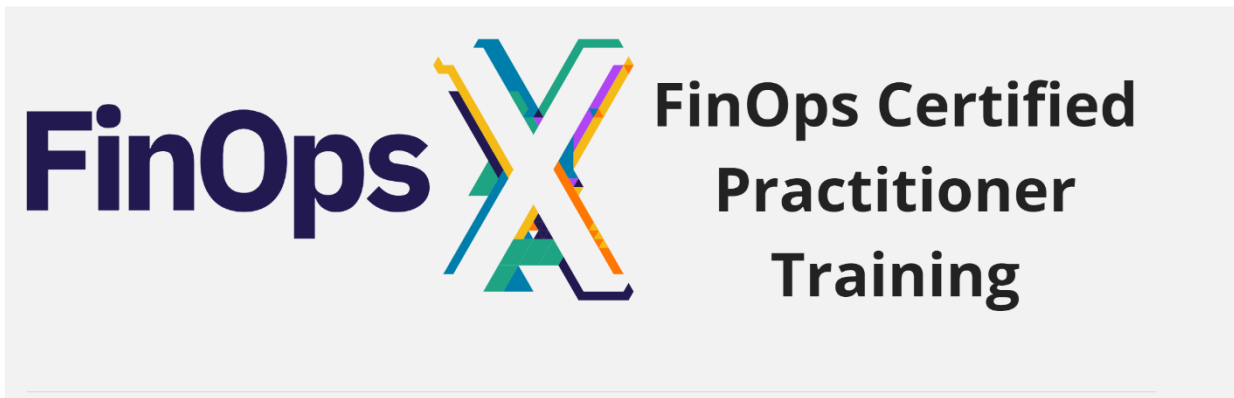
Canada Revenue Agency – FinOps Training

When CRA began its FinOps journey, they wanted to provide the opportunity to stakeholders from technology and finance to understand FinOps practices. The FinOps Foundation is the learning provider they turned to for increasing organizational learning

Statistics Canada – FinOps Reporting

To enable departmental monitoring and governance over cloud spend, StatCan has invested in tailored dashboards and reporting to attribute costs back to lines of business and fund centres.

---

[i] Delivery Team refers to a multi-disciplinary team responsible for delivering a product using cloud. Deliver Team is used to denote the shift from organizing as functional silos towards multi-disciplinary teams made up of developers who build and run the product, architects, security, product managers, and user interface experts.