Managing Op Ex in A Cloud Data Platform

One of the key selling points of building data platforms in the cloud is it eliminates upfront capital expenditure. Since it is a pay as you go approach, the cost model transitions to an operational expenditure model.

While the operational cost of a cloud infrastructure would not be significant at the outset; it could grow unexpectedly as more data platforms are migrated into or built in the cloud. As a result, operational expenditure could grow out of control if an organisation's cloud transformation program is not well executed.

There are some key considerations to reigning in the operating expenditure of data platforms in your cloud:

- Chargeback model
- Architecture and design of data platforms
- Development of data platforms
- Operational concerns

Chargeback model

When an IT division operates as a profit center, the cost of operating an infrastructure could be apportioned to the business units based on utilisation. The assumption is there is an effective way to track utilisation whether it is a single tenancy or multi tenancy model across data platforms. The cost is then charged back to the business units and is budgeted for by that unit.

This is a good option for managing costs of operating cloud platforms and validates the business case for a data platform. As the business users leverage the data platform, it could grow organically to more use cases. IT could also expand the data platform into other business units. Conversely, if the data platform does not deliver sufficient business value, it will be wound down or repurposed.

The chargeback model also drives accountability within IT organisations to track utilisation, adhering to SLA and managing capacity. Utilisation could be tracked across multiple measures like:

- Compute which could be dedicated clusters, DBUs, DTUs or vCores, etc.
- Storage space in Data Lakes, Databases, Data Warehouses, or other forms of storage
- Ingress and egress

Architecture and design of data platforms

The architecture and design of new data platforms must match functional needs and be flexible. This is especially the case if it is for an organisational wide data platform. Careful consideration around fit to business case, data, storage, network, security, and system components among others is required.

The "build it and they will come" approach will not work in all situations. It could lead to the implementation being wound back, or the products/components being replaced when business value is not realised, leading to more implementation costs due to rework.

Decisions around products to use for a solution must fit the business case rather than a grand narrative. Architects attempting to future proof the data platform or fit obscure requirements would need to strike a balance. Decisions made could lead to implementing products or components that ill-fit the requirement, are overly complex or costly to run. In some cases, business process reengineering rather than a technical change could be the answer.

Keeping an open mind about how the system architecture could evolve over time is critical. Services or components that are no longer relevant or fit for purpose could be swapped out, thus reducing costs. Similarly, system architecture could evolve as prevailing technology and approaches change over time. If the architecture and design is loosely coupled and well thought out, it would be simpler to decommission older components and integrate-in new components. As we know, optimisation exercises are continuous.

The approach of migrating data platforms to the cloud rapidly using IAAS platform and dealing with cloud specific redesign later can be problematic. While this works at the beginning, especially in the case of a distributed user base and/or an aging infrastructure, it pushes issues further down the road and leads to higher operational costs over time until the platform is re-architected to take advantage of cloud platform capabilities. It is best to devise a staged migration plan which incorporates redesigning and reengineering data platforms.

Development of data platforms

While development of the data platforms is strictly not an operational cost, it could fall into the operational expense bucket because the development work is usually ongoing. Approaches listed below would help mitigate the cost of running a development team:

- Good program management and project management practices backed by good tools that flow into DevOps processes. This relates to a macro and micro view of the program of work
- Test driven development is not well adopted in the data world but is becoming more important. Upfront validation within limited scenarios would ensure less rework as the project flows into UAT (User Acceptance Testing) and is rolled out into Production
- Strong Dev Ops Culture with reliable source control processes (code merge) and automated deployment of IAC (Infrastructure as Code) and Data Platform Code is key.
- Adoption of Data Contracts by data publishers. This is a new concept akin to well-defined APIs for communication used by microservices
- Deliberation over data movement cost as volume of data movement determines utilisation of compute, network bandwidth and lastly storage. There needs to be thought put into data loading mechanism as in incremental vs. full vs. CDC (change data capture). There could also

be chatty calls into REST APIs when ingesting data and the regularity of these calls could be revisited.

Operational concerns

Support Team

Operating expenses include the support teams that keep the cloud environment running. A skilled operational team is required to focus on deploying, managing, and monitoring the assets. Another key function is to optimise the resources to fit business needs.

Efficiency and effectiveness are key measures to managing the operational support cost. Thus, it ties back to the architecture, design, and development considerations:

- Architecting and designing data platforms to be simple and maintainable
- Standards around system architecture. For example, the more standardized the data flows are, the simpler they are to maintain
- Well defined operational architecture. automation for alerting based on monitoring of resources which then flow into support ticketing systems.

There should also be policies around when assets should be shut down and cleaned up after a Proof of Concept or Project is completed. Similarly, data that has fallen into disuse could be archived rather than stored in premium storage. Based on data governance rules, some confidential data could also be purged.

Matching The Capacity Needs To SLA

A key factor around operating costs is right sizing your infrastructure to meet the needs of the business. On the ingestion and data transformation side, the volume of transactions and data as well as time window sensitivities determine the capacity needs. On the analytics side, the number of users, number of reports and complexity of analytics drive the capacity needs. As for data science, the number of models and their complexity determines the capacity needs.

Generally speaking, compute and storage are decoupled in the cloud. Scalability of the infrastructure depends on the number of cores, amount of memory, storage performance and data distribution strategies. Given that it is the cloud, elasticity can be leveraged to manage costs.

There is also a consideration around whether some capabilities in the infrastructure should run 24/7 or run only within a shorter time window. There are some services that could be started up when the endpoint is connected to by users. There would be a brief time delay while the service starts up and there are ways to keep the service alive during busy periods.

Storage pricing is driven by performance tiers and in some cases disks which have differing I/O characteristics. Some rework is required for the case of storage or data distribution if there are changes. An example is when data is repartitioned to suit common query patterns.

It is always good to review cloud real estate and understand where there is more compute and/or storage capacity than required and where more should be provided. There are native capabilities for monitoring and alerting around capacity utilization.

Distributed Infrastructure for Geographically Distributed User Base

When there is a geographically distributed user base, there is always a tension between centralizing the assets vs distributing them. If the locations are across geographical regions, reporting capabilities for example could be regionally co-located. Due to considerations like customer satisfaction, cost, and complexity, some amount of balancing of the needs is required here.

The same considerations apply to data as well. Questions would be whether a subset of data or summarised dataset should be replicated as opposed to the full dataset. An example is different sites hosting local data at the closest cloud location and data movement for consolidation of data at corporate level. Some of the products or components might natively support moving data across locations.

Conclusion

As described above, management of operating expenses in the cloud is multi-layered and cuts across different areas of concern. It would be ideal to have an overall program or an organisational wide approach to be successful at managing the costs of running cloud environments.

If your business needs help with managing their operating expenses in the cloud, or you would like a review, please contact our Fujitsu Data & AI specialist now.

Note: For the sake of brevity, I did not cover areas like High Availability and Disaster Recovery. Some cloud providers do provide these capabilities at different price points for redundancy across zones within data centers, data centers or regions.

Contact

Fujitsu Data & AI

+61 3 9924 3000