# Architecture Framework Advisory Committee (AFAC)
# Enabling the Government of Canada's Enterprise Data Strategy

## Discussion Summary

June 14, 2019 (13:00 to 16:00 EDT)

### Event purpose statement

In support of the *Policy on Information Management*, the Government of Canada (GC) is looking to enable an Enterprise Data Strategy, supporting technologies and processes within a shared infrastructure environment; however, it also recognizes the need to transform GC architecture to support this new reality, manage its downstream effects, and support new services for Canadians.

### Highlights

The key is to start with a limited number of use cases on which to the organization builds on. This approach allows an organization to map its current capabilities leveraging defined use cases and help define a starting point to progress upon. It is advised that an organization should not restrict or cement its data strategy solely upon data lake platforms as many other data management platforms are available and can be better suited to some use cases.

Standardization of data is key to faster benefit realization. Outlining data governance from the very beginning is crucial. It is important this is kept broad in areas to allow for flexibility.

Data Privacy and Security should be considered from inception and needs undergo constant review and improvement. Being situational and reactive allows organizations to be application driven and avoid unnecessary effort.

The sensitivity of the data needs to be considered, but also the result of compute on that data and its cross reference with other data sources

Organizations need to invest in data literacy for their workforce and attract talent in data analysis and data engineering. A 2:1 ratio regarding data analysts to data engineers is recommended.

### Key Considerations

When adopting a data a management strategy/ platform the first step in the process is defining a clear definition of the business problem. This involves verifying the outcomes and use cases the customer is trying to achieve. From there, a model can be constructed incrementally. This involves starting small and a scaling up the solution. The intent of the data should be established, as well as the risk associated with it. The goal should be the adoption of a single data strategy/platform. However, this is not always feasible and can turn into a data pipeline which is why it is important to define data lineage and where to place the platform within your enterprise ecosystem. Doing so, will require data governance which should be broad enough to allow for flexibility. It is important to quantifiably measure efficiency improvements throughout the implementation and evolution of an enterprise date strategy.

From a security and privacy lens there is a strong need for defined principles and standards. An organization should assess their security requirements on a case by case basis allowing for the correlation of datasets, and data contained within, against appropriate security levels.

An organization should analyze use cases and strive to be application driven. Pitfalls in efficiency are often attributed to the fact that 90% of data should have the option of being sharable but due to poor data governance and over-classification it is not. This further underscores the importance of properly analyzing the data and ensuring that the level of security is justified.

Systems should be built to manage data governance (i.e. Governance should be executable). It is unwise to lock into a specific data management solution (data lakes). The solution should be selected based on the specific use cases it satisfies. Closed systems should be avoided in this regard, where they do not maintain flexibility. The system should require authentication.

Catalogue automation is an important factor when avoiding a "data swamp". It is crucial to understand the domain, and what to integrate. Large scale automation can be difficult depending on the type of data due in part to machine learning algorithms' struggle when cataloguing data. Cataloguing is about automating the interrogation process of the data. Careful separation of concerns is a key factor. It is strongly advised that standards be implemented but kept basic and simple at first as this allows an organization to ease into the process. Having an abstraction layer can help map the data in lieu of industry standards.

There is wide consensus that bringing together structured and unstructured data can maximize value for an organization.  However, there are tangible challenges to doing so due in part to the differences of approaches when it comes to mapping data. Transforming unstructured to structured data remains the best way of handling unstructured data. This remains a difficult process but as technology progresses it is anticipated that this process will become easier to accomplish.

To enable the GC to architect a proper data management strategy certain skillsets are required. It is strongly advised to outline ones contextual needs at the outset as this will drive your recruitment and personnel training strategy.  Having data oriented expertise on hand that understand your data domain and models is instrumental as it ensures the right data is being fed into the proper siloes. The divides between technical and non-technical individuals need to be broken. The focus should remain on data engineers and data scientists, but non-technical/non-traditional roles need to be on-boarded as well. In turn, a multidimensional view is given to the project. Some of these non-traditional roles include outreach personnel. These individuals are responsible for polling the community to see what is currently being done with data and what could potentially be done with data. Change management roles are also deemed crucial for this process. When an organization can verify with experts what its capabilities are given proper access to different data sets this drives better application development.

The GC needs to select a data strategy that provides flexibility to ensure that new technology can easily be introduced. After selection, it is important to heed the data strategy. With the development of platforms having many viable products, this gives the GC flexibility. New services should be integrated on a smaller scale first and grow thereafter. This approach allows an organization to gain momentum and at the end of this process service level agreements can be upheld much easier. It is important to be proactive when offering services, client outreach in anticipation of service issues drives customer satisfaction.