CPA CHARTERED PROFESSIONAL ACCOUNTANTS CANADA

Inject Trust into Your Data Sharing Ecosystem

August 2021

CPA Canada Foresight
REIMAGINING THE PROFESSION.

21-2078

ABOUT CPA CANADA

Chartered Professional Accountants of Canada (CPA Canada) works collaboratively with the provincial, territorial and Bermudian CPA bodies, as it represents the Canadian accounting profession, both nationally and internationally. This collaboration allows the Canadian profession to champion best practices that benefit business and society, as well as prepare its members for an ever-evolving operating environment featuring unprecedented change. Representing more than 220,000 members, CPA Canada is one of the largest national accounting bodies worldwide. cpacanada.ca

Electronic access to this report can be obtained at **cpacanada.ca**.

Table of Contents

This primer focuses on data sharing ecosystems and on controls that will generate trust. CPAs are well positioned to help design and manage controls that will bring clarity and transparency to data, sharing platforms and artificial intelligence (AI). Systems and approaches are evolving quickly but no standards or detailed guidance exist yet to maintain appropriate controls on data sharing ecosystems. CPAs' technical training and proven competencies in managing financial information controls can and should be transferred to data value chains in order to generate much-needed trust.

Managing a growing trust deficit

The biggest obstacle most organizations face in their efforts to achieve digital transformation is usually not technical; it's maintaining trust across data value chains. If you work upstream, where data is generated and collected, allowing data to be shared and reused by others outside of your span of control requires a great deal of trust. And if you work downstream in data science, trusting that the data you receive is accurate, complete and up to date is paramount for success. As the number of use cases and applications grow in size and in scope, ecosystems become more complex. Trust needs to be distributed among a larger number of business units, data-collection intermediaries and data scientists/engineers.

In order to inject trust into the data sharing ecosystem, your organization needs credible systems and controls. Data will be housed in different locations, from cloud-based platforms and specialized third-party data lakes to your own suite of local servers and edge computing devices. Nevertheless, you will need to ensure that *only the right individuals with the right credentials can access the right data at the right time*. Organizations need to address inter-related challenges head on in order to operate trustworthy algorithms, demonstrate compliance to privacy regulations, keep customers satisfied and help avoid costly data breaches

Digital transformation

A well-functioning data sharing ecosystem is essential to support your organization's digital transformation. Ideally, it is based on solid data governance rules. An important initial step is to adopt a corporate data policy. Other helpful actions include the approval of a digitization strategy along with a budget, and the creation of a hybrid team composed of subject matter experts, relevant data experts and IT. To be successful, the team would begin by identifying specific business problems to solve, or business opportunities to exploit, then creating and documenting use cases and selecting appropriate AI models to train. When these initial steps have been taken, organizations benefit from focusing attention on creating high quality data. Data needs to be clean, accurate, complete and annotated properly. This can be accomplished through data collection and preparation activities. Once datasets are prepared, they can be made available for sharing with your hybrid team.

Increased reliance on cloud-based platforms

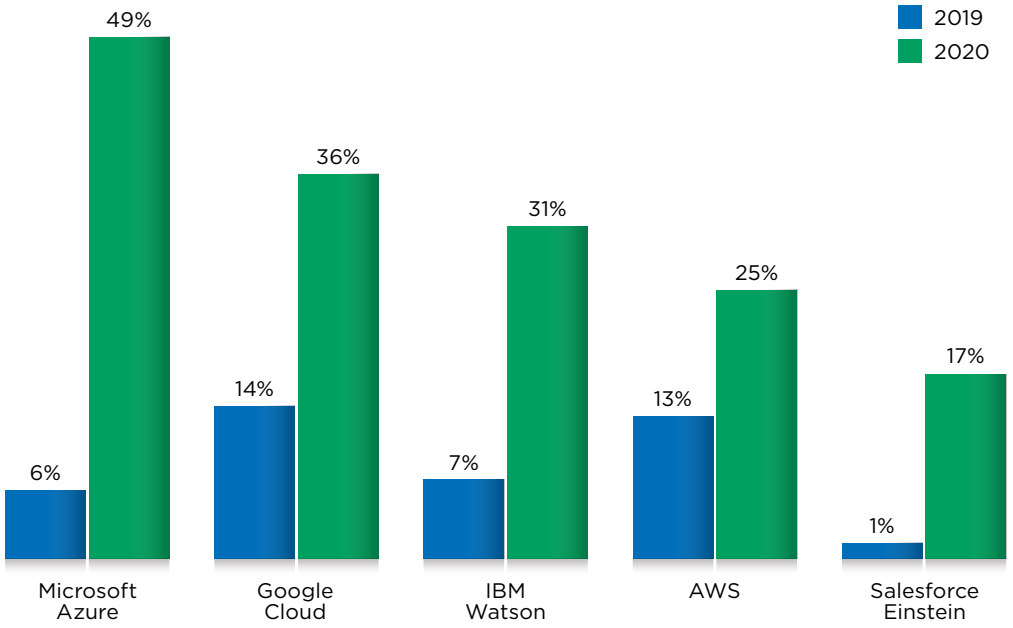
A few years ago, organizations generally initiated their digital transformation journeys by looking for suitable datasets and experimenting in-house with AI. Managing data access was relatively straightforward, as data was generally housed in local servers. But one year into the COVID-19 pandemic, it became clear that organizations were forced to adopt hybrid approaches. Many needed support to expand memory, help a high volume of users, update machine learning models to reflect the new normal and sustain their workforce online. This explains the meteoric rise in popularity of online machine learning platforms in 2020. The rate of adoption is such that leading research and consulting firm Gartner predicts that public cloud services will be essential for 90 per cent of data and analytics innovation by 2022.[1]

There are clear advantages for organizations to use cloud-based platforms. Many of the required tools, software and hardware needed for data preparation, model training and deployment are included in subscription costs. When training models, you only pay for what you need in terms of CPU computing time. A growing number of organizations recognize that global machine learning platforms are also better equipped than legacy systems to deal with other issues such as cybersecurity, IT modernization and data access.[2]

1 Gartner, 2020. Top 10 Trends in Data and Analytics for 2020. https://www.gartner.com/smarterwithgartner/gartner-top-10-trends-in-data-and-analytics-for-2020/

2 Karthik Ramachandran and David Linthicum, 2020. Why organizations are moving to the cloud. Deloitte: January 2020. https://www2.deloitte.com/us/en/insights/industry/technology/why-organizations-are-moving-to-the-cloud.html

Figure 1: Data science and machine learning tools/framework used by early adopters

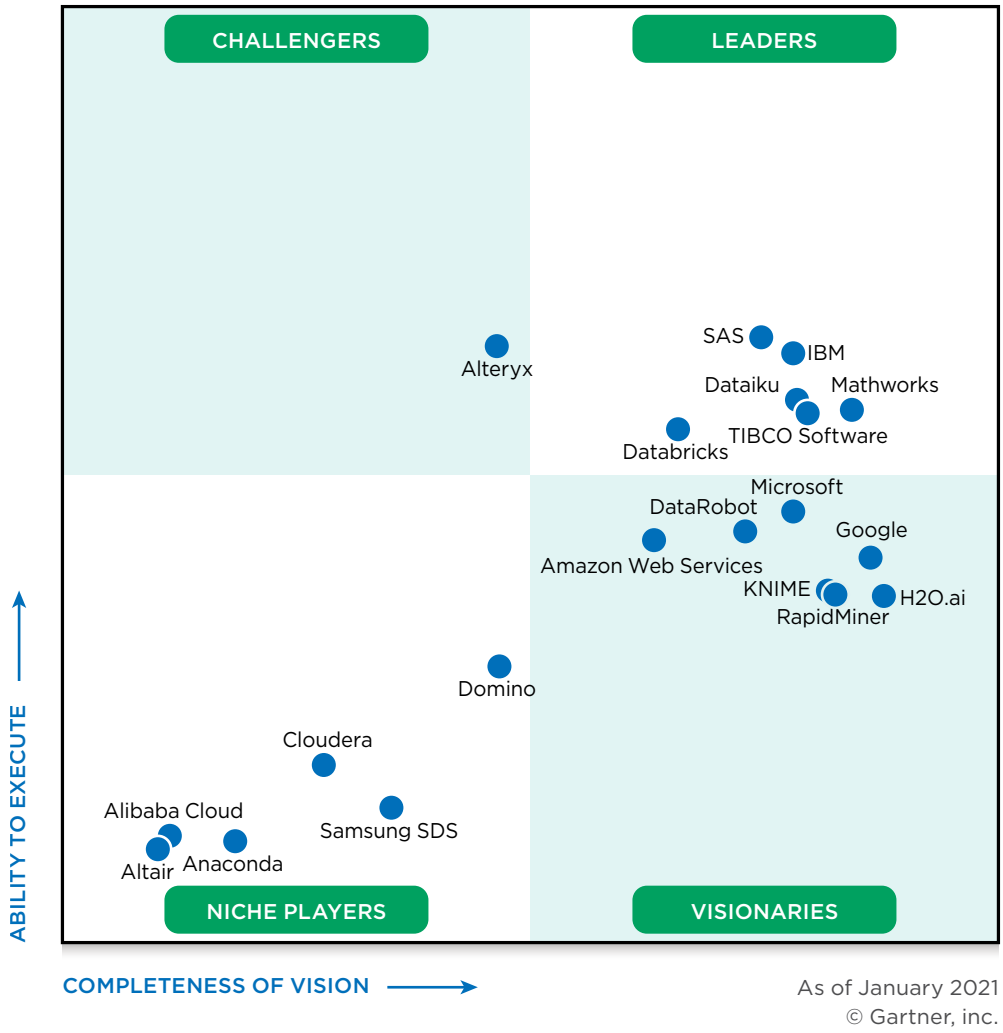


Source: Appen, 2020. The State of AI and Machine Learning. https://resources.appen.com/wp-content/uploads/2020/06/Whitepaper-State-of-Ai-2020-Final.pdf

Gartner recently provided an evaluation of 20 machine learning platforms that can be used to source data, build models and operationalize machine learning. Through these platforms, organizations can literally do it all and scale when ready. Raw data can be stored, prepared, annotated and labeled, and appropriate algorithms can be selected, including explainable artificial intelligence (XAI). Online platforms offer vast computing power that is required to train machine learning models with your data. They can also offer additional datasets if necessary and provide support for model deployment, monitoring and maintenance.[3] As the graph below shows, the industry leaders include organizations such as SAS, IBM, Dataiku, MathWorks, TIBCO Software and Databricks.

3 Gartner, 2021. Magic Quadrant for Data Science and Machine Learning Platforms. Gartner, March 1, 2021. https://content.dataiku.com/nurturing-why-ai-platforms-lob/gartner-mq-21?utm_campaign=CONTENT+Nurturing+Workflows+June+2020&utm_medium=email&_hsmi=90911789&_hsenc=p2ANqtz-8PKZwucx5qQZoOZvnCh8uiFqEL3U-_IqFJiH_m8Icw-M7V2x1dT50UjgzI-ixPZKQ7SzaCj7vYLS2u4_fK8iBXMJJT3A&utm_content=90911789&utm_source=hs_automation

Figure 2: Magic Quadrant for Data Science and Machine Learning Platforms



Source: Gartner, Magic Quadrant for Data Science and Machine Learning Platforms, March 2021

Interestingly, organizations are not putting all of their data in one basket. Most report using more than one service provider as they broaden the deployment of AI and machine learning tools to various business functions. Customers want to remain independent and avoid locking in with one vendor as they leverage major cloud providers.[4] Proprietary software is used to bridge datasets, cloud applications and applications managed by specialized third parties as well as private data centers.[5]

4 Dataiku, 2021. Getting the Most of AI in 2021. Dataiku, January 2021, p. 14.

5 Dataiku, 2020. 2021 Trends: Where Enterprise AI is Headed Next. Dataiku, 2020. https://www.dataiku.com/stories/2021-trends-where-enterprise-ai-is-headed-next/

Although platform service providers feature tasks and processes related to data access and controls (such as data and governance models and collaboration), organizations are expected to design, implement and monitor data sharing and access policies to manage data access.

According to Tari Dwiek, director of Technology Alliances at Snowflake (a data science and AI firm), organizations want secure and governed access to data. "Now that the cloud is opening up the ability to manage data at scale, customers see both the opportunity and the critical need to enable data governance at scale."[6]

For organizations getting more sophisticated at using AI and machine learning, the trend is to democratize data use. This requires allowing a larger number of analysts and subject matter experts to access data. New service offerings are being introduced to manage data through "data lake" architecture. A data lake is a repository where raw data can be stored without preparation. Client organizations can store and use business operations data on data lakes while allowing analytics work to be performed live, without having to copy, label or annotate data.[7]

Increasingly, data sharing is also occurring between organizations. Data sharing protocols and protection requirements also need to cover suppliers and vendors. As such, terms and conditions of existing contracts need to be updated to reflect data policies, procedures and protocols. Issues such as data ownership, IP and copyright, data residency requirements and adherence to relevant privacy and ethics rules must be addressed.

6 Dataiku, 2021. Getting the Most of AI in 2021. Dataiku, January 2021, p. 14.

7 Dremio, 2021. The Next-Generaton Cloud Data Lake: An Open, No-Copy Data Architecture, 2021. https://hello.dremio.com/wp-the-next-generation-cloud-data-lake.html

CPAs should manage new data sharing controls

Given these hybrid scenarios involving multiple providers, data sources, models and outputs, managing data access and reporting on compliance has become paramount. In order to generate much needed trust to entice various units to share high quality data for future reuse, there is a need for a credible accountability framework. In the context of digital transformation, a data controller is responsible for the stewardship of data shared for the purpose of data reuse and to enhance the value of data through its protection, curation and appropriate usage.

Traditionally, controllers protect resources and ensure that only people with the appropriate access rights (need and permission) are authorized to use the resources. Professional accountants have traditionally performed stewardship roles in relation to financial and physical resources. In this capacity, they ensure that the financial resources of the organization are protected, that related laws are adhered to and that activities undertaken by the organization are strategically aligned.

A data controller's stewardship role is not limited to financial resources. It applies to all data resources and will overlap the financial controller's role. It is a natural evolution for professional accountants to broaden their traditional financial stewardship roles to include all data slated for data sharing and reuse. However, stewardship does not mean ownership. Non-financial data owners will often be the line functions within an organization. Stewardship is an enabling function to ensure that the data owners protect, curate, share and use the data according to external (laws, regulations, etc.) and internal (policy) constraints.

As markets are created for the exchange of data between buyers and sellers, it is expected that formal requirements will be developed to certify the data sold, shared or traded. The certification will likely entail the ability to prove the accuracy and source of the data, also known as provenance or lineage. A data controller would likely be called upon to provide this certification.

Organizations have significant opportunities to use new technologies to their advantage, but these opportunities raise equally significant legal and ethical challenges. Not all uses of technology align with the values of different societies with respect to fairness, security, privacy, understandability and transparency. As stated in the International Code of Ethics for Professional Accountants (the Code), taking into account their position and seniority in the organization, professional accountants are expected to encourage and promote an ethics-based culture in the organization – and as such are well-positioned to help organizations in a data controller role.

The key element of this role is to protect the data. The first layer of data protection is to ensure that only authorized individuals have access to the data. Another layer is to ensure that the origin of the data can be demonstrated. Stewardship also means ensuring that data is used for its intended purpose.

An enhanced role for professional accountants is to ensure that jurisdictional boundaries are respected through appropriate monitoring of data usage. An equally important role is to certify that the data being used or sold is fit-for-purpose in that the lineage and provenance of the data can be proven.[8]

8 CPA Canada and IFAC, 2021. Professional Accountants' Role in Data: Discussion Paper. Joint Paper. CPA Canada and IFAC, January 2021. 21 pages.

Time for MLOps?

Managing a data sharing ecosystem's performance and reporting on compliance increases in complexity as organizations democratize data reuse. When digital transformation processes are being systematically embedded in daily operations, the small hybrid team's model can evolve and expand to manage what is called Machine Learning Operations or MLOps. This new discipline emerged in late 2018 in order to help organizations scale (go from managing just one model to managing hundreds or thousands), manage complexity and mitigate risk introduced by the use of machine learning models. MLOps relies on the expertise of subject matter experts, data scientists, data engineers, software engineers, DeVOps (software development engineers), machine learning architects and model risk managers and auditors. It aims to standardize and streamline machine learning life cycle management.

In a recent O'Reilly publication on MLOps, Mark Treveil from Dataiku argues that strong data governance processes and controls are essential elements of successful MLOps transitions. Decisions and outputs need to be tracked and documented, and controls need to be put in place in order to answer the following questions for every project:

- What is the data provenance?
- How was the original data collected and under what terms of use?
- Is the data accurate and up to date?
- Is there private information or other forms of sensitive information that should not be used?[9]

Through MLOps, organizations aim to "have an overall view of which teams are using what data, how and in which models." This "also includes the need for trust that data is reliable and being collected in accordance with regulations as well as a centralized understanding of which models are used for what business processes."[10] That being said, there are currently no clear methodologies in place to monitor and report on data governance compliance. Specialized firms have begun to introduce new platforms and software to

9 Mark Treveil and the Dataiku team, 2020. Introducing MLOps: How to Scale Machine Learning in the Enterprise. O'Reilly, November 2021, 169 pages.

10 *Ibid*, p. 10.

manage data access governance. As an example, Immuta and Databricks now provide tools to automate and control data access to support MLOps teams. These tools use processes such as Attribute-Based Access Controls (ABAC), automated security and privacy controls such as de-identification and sensitive data discovery and tagging, and data access auditing and reporting.[11]

With appropriate governance mechanisms, an organization can also ensure that it delivers on its responsibilities to all stakeholders and abides to the fundamental principle of fairness.

11 Immuta and Databricks, 2020. A Guide to Data Access Governance. Immuta and Databricks, 2020. https://www.immuta.com/downloads/a-guide-to-data-access-governance-with-immuta-databricks/

Looking forward

This article points to the need for controls to inject much needed trust into data sharing ecosystems. The need for and importance of trust has grown significantly in an increasingly digitized world. When it comes to financial reports, oversight and controls, CPAs represent the gold standard of trust. As a profession, CPAs are poised to lead the conversation on trust across data value chains – from data collection, sharing and access to trust in AI and machine learning tools. By monitoring, verifying, auditing and reporting, along with auditing compliance to best practices on data sharing, access and reuse activities taking place in multiple locations and formats, CPAs can help demonstrate that new systems are operating as intended.

CHARTERED
PROFESSIONAL
ACCOUNTANTS
CANADA

277 WELLINGTON STREET WEST
TORONTO, ON CANADA M5V 3H2
T. 416 977.3222 F. 416 977.8585
CPACANADA.CA