Data Quality and Data Availability

February 2024

AI has raised the stakes for existing challenges to getting the right data, at the right time, and with consistent access across borders. It has also amplified issues of accuracy, completeness, reliability, and relevance. As Generative AI (GenAI) and Large Language Models (LLMs) have become a new focus for development, these issues in quality, availability, and use of data are becoming more consequential. In this edition of DataTalk, we discussed data quality and data availability trends, and how they are shaping the trajectory that algorithms will take in the future. This briefing note summarizes the discussion held on February 27, 2024, respecting that the forum is conducted under the Chatham House Rule and does not represent the official position of the IIF or its membership.

Accurate outcomes based on trusted sources. The financial industry has designed robust data management processes and practices around different pillars that support the quality of the data they use for various purposes. These procedures include considerations that are two-fold: (1) data movement between systems and data usage –to track what data is being used, and by whom–; and (2) data privacy and data storage –to ensure confidentiality, safety, integrity, and compliance–. These elements of data management allow Financial Institutions (FIs) to then build AI and GenAI use cases on top, ensuring they can provide accurate outcomes based on trusted sources.

Complex by nature. On top of companies' internal processes, FIs also need to overlay the regulatory requirements put in place in the countries or regions where they operate and establish permissions on the different data stacks depending on where they can flow (or not flow) across different jurisdictions and frameworks. This depends on different sets of regulations, some of which establish data localization requirements or outright prohibitions to transfer data across borders.

Trust by design. Companies need to be confident about the foundations of their data, features such as data quality, availability, metadata, and the governance of data need to be robust before a firm can put data into new tools like GenAI and LLMs. To this end, some FIs are putting controls and policy frameworks in machine-readable formats so that the changes those frameworks undergo can be introduced into a system that ensures compliance over time, which then becomes a living part of the data management lifecycle. Firms that fail to apply these or similar measures or controls, could end up spending more time and resources designing, testing, and discarding models based on the impossibility of having GenAI models and LLMs unlearn things they have seen in the data used to train them.

Multiple actors across the data and AI value chain. There are multiple parties involved in the use of data for financial services: global authorities, financial regulators, FIs, third and fourth parties, and others. In this context, building harmony (through standardized formats, regulatory or supervisory equivalents, and cultural changes within firms) is paramount. Failing to achieve this harmony among the various actors and jurisdictions involved negatively affects data quality and data availability, stifles innovation, hinders market integrity and financial stability, and diminishes the chances of getting ahead of bad actors in key areas such as AML/CFT, fraud prevention, resiliency and more.

*We look forward to continuing the DataTalk series on Wednesday, May 22, 8:00 am Washington DC / 8:00 pm Singapore when we will explore BigTechs: Business Models and Platforms. Our new format of this event gathers thought leaders and experts sharing their insights and views, as well as a dialogue with member firms about implications for the financial services industry and new developments on the policy front. If you would like to register for the upcoming session of DataTalk, please **register here**.*