THE "DATASTACK": A DATA AND TECH BLUEPRINT FOR FINANCIAL SUPERVISION, INNOVATION, AND THE DATA COMMONS

BFA GLOBAL

R2A

REGTECH FOR REGULATORS ACCELERATOR

Simone di Castri,
Matt Grasser, and
Arend Kulenkampff

May 2020

www.bfaglobal.com/project/R2A

@BFAGlobal
@R2Accelerator

Table of Contents

Contents

Executive summary

Digital transformation of the financial sector has been driving financial inclusion and innovation.[1] Better connectivity, cheaper mobile technology, new products, and improved customer experience are steadily closing gaps in technology adoption and financial access.[2] However, a new "Big Data divide" is forming between the digital financial services (DFS) market participants who can leverage data for their respective ends, and a number of public and private sector stakeholders who cannot. This imbalance of power has the potential to skew the development of digital financial ecosystems in favor of a few large players.

To bridge this divide, a new Open Data Commons needs to be established that guarantees access to rich and plentiful data, as well as the tools to make them meaningful and actionable. This report proposes an expanded variety of Open Data Commons, which we refer to as DataStack. **The DataStack is a modular, streamlined, end-to-end data architecture that leverages an interoperable data platform and advanced analytics tools to generate meaningful, actionable insights in digestible formats for multiple personas.**

This vision and blueprint presented in this report are the outcome of three projects that the BFA Global team undertook from 2016 to 2019. These projects were designed to accelerate the adoption of advanced data analysis methodologies and modern tech development process by the public authorities that regulate and supervise the financial sector:

1. The RegTech for Regulators Accelerator (R^2A), implemented in Mexico and the Philippines in partnership with the Bangko Sentral ng Pilipinas (BSP) and the Mexican Comisión Nacional Bancaria y de Valores (CNBV), and sponsored by the Bill & Melinda Gates Foundation, the Omidyar Network, and USAID.[3]

2. The "Data Stack", implemented in Nigeria in partnership with the Central Bank of Nigeria (CBN) and the Nigeria Inter-bank Settlement System Plc (NIBSS), co-sponsored by the Bill & Melinda Gates Foundation.[4]

3. The "Gender disaggregated data and women financial inclusion diagnostic" implemented in Egypt in partnership with the Central Bank of Egypt.[5]

Through this body of work, we came to realize that (i) there is a "data divide" that threatens to favor a few players, reduce competition, and undermine innovation; (ii) an Open Data Commons would address these challenges and risks; (iii) an Open Data Commons requires a new approach to building and structuring regulatory/supervisory data architectures; and (iv) a core DataStack centered around financial authorities and government agencies would provide multiple users a modular, interoperable Open Data Commons.

1 GSMA (2019), State of the Industry Report on Mobile Money.

2 GSMA (2019), The Mobile Economy.

3 Simone di Castri, Matt Grasser, and Arend Kulenkampff (2018a), Financial Authorities in the Era of Data Abundance: RegTech for Regulators and SupTech Solutions, BFA Global RegTech for Regulators (R2A) white paper. See also: www.bfaglobal.com/R2A.

4 www.bfaglobal.com/insights/datastack.

5 www.bfaglobal.com/insights/Egyptgenderdata.

A vision of an Open Data Commons to close the Big Data divide

The emerging Big Data divide

There is an imbalance of power over the control of data and command of data science that has the potential to skew the development of digital financial ecosystems in favor of a few large players. On one side of the "Big Data divide" are financial services and technology providers that can successfully harness data from their customers and the ecosystem. They possess the scale and capital to accumulate vast stores of data from which to extract value using the latest advances in data science. On the other side of the equation are the data "have-nots" who are unable to access or absorb the surfeit of data being generated by the digital financial ecosystem, and/or who lack the tools to reap the data divide.

Data assets, data infrastructure, and data scientists are costly investments for providers to either acquire or develop within their companies. These can act as barriers to entry. Indeed, the cost of financial and market data has grown rapidly in recent years, with some polls suggesting that data is the largest expense item for early stage startups.[6] Furthermore, the mastery of Big Data and artificial intelligence (AI) in business decision-making can exacerbate powerful competitive advantages, and those with access to huge customer networks in other lines of business, such as China's Ant Financial or e-commerce giant Amazon, also benefit from network effects. Ultimately, the data-rich and tech-savvy may come to squeeze out the smaller players and potential upstarts, to the detriment of overall market competition. The evolution of the broad technology sector, which has come to be dominated by a handful of superplatforms, is a worrying portent for digital financial services (DFS).

The Big Data divide also poses risks to consumers. On one hand, they benefit from improved access, customer experience and choice as a result of the application of Big Data and AI in areas such as credit risk management. At the same time, consumers often lack visibility into how their data is harvested and exploited. As financial service providers (FSPs) gather more and more data about their customers, including from social media and the Internet of things (IOT), the information asymmetries that underlie traditional financial intermediation may flip in favor of the former. That is, AI enables providers to capture more information about the financial lives of their customers than they might know about themselves.[7] This asymmetry could prompt some providers to target their offerings to low-risk or data-rich market segments to the exclusion of high-risk or data-poor populations. The Big Data divide could therefore give rise to new forms of redlining and "cream skimming,"[8] leading to new patterns of financial exclusion - even as fintech opens

6 See Burton-Taylor (2019), Global Spend on Financial Market Data & News Topped $30b for the First Time, Strongest Growth Since 2008. And Dough Nelson (2015), The FinTech Paradox in the Age of Open Data, Medium

7 For instance, credit risk scoring models using machine learning techniques might make lending decisions based on correlations between social-media profiles and web-browsing activity that are undetectable by humans. See: Economist (February 2017), Big data, financial services and privacy: Should our bankers and insurers be our Facebook friends?

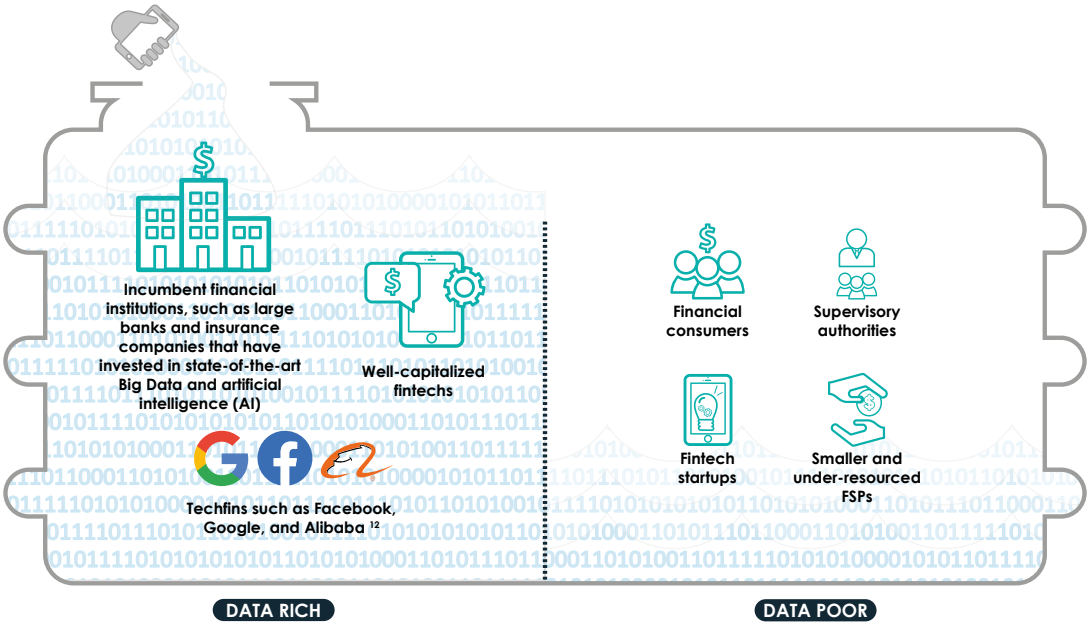
8 Cream skimming refers to the business practice of providing a product or a service to only the high-value or low-cost customers of that product or service, while disregarding clients that are less profitable for the company.

pathways to inclusion.[9] Besides concerns around data privacy and consumer protection, the profusion of products and providers can also be overwhelming for customers. Without independent and digestible information about firms' records of conduct and their product offerings, information and choice overload[10] may make customers even more susceptible to misuse and abuse.

One obvious way to guard against the harmful side effects of information asymmetries is regulation. Financial authorities have licensed many new Big Data and AI-based business models, and defined the rules of the game for new entrants to the financial ecosystem. However, in their supervisory capacity the authorities often lack the requisite data management and data science to perform their oversight tasks properly. Thus, for example, the outputs of machine-learning (ML)-driven credit risk models that draw on vast quantities of data are often difficult to interpret. Without adequate data infrastructure and know-how in place to properly validate these models and scrutinize their results, regulators may opt to throttle the pace of financial innovation rather than risk under-regulating sector.

Financial sector supervisors face many challenges. They have access to a growing wealth of data from new and existing sources to guide their policy and decision-making, yet they often lack the infrastructure or skills to absorb the windfall, as surveys conducted by the BFA Global RegTech for Regulators Accelerator (R^2A) have shown.[11] Rather, more data simply translates into more manual data processing, leading to "analysis paralysis" down the line. "Black box" models add to the complexity of supervision of new products and players.

FIGURE 1. THE BIG DATA DIVIDE



9 Majid Bazarbash (2019), FinTech in Financial Inclusion: Machine Learning Applications in Assessing Credit Risk, IMF Working Paper 19/109.

10 Choice overload refers to the paradox that people exhibit difficulties making decisions when faced with many options.

11 BFA Global RegTech for Regulators Accelerator (R^2A) (2019), The State of RegTech: The Rising Demand for "Superpowers."

12 See: Zen Hoo (December 2, 2016) "TechFin: Jack Ma coins term to set Alipay's goal to give emerging markets access to capital," South China Morning Post.

Open Data Portals help to bridge the data divide, but often fall short

To counteract the emerging Big Data divide, a more level playing field needs to be established in terms of both access to and usage of data. Open Data Portals are one such leveler. Data are open when "data and content can be freely used, modified, and shared by anyone for any purpose."[13] Open Data Portals enable access to open data, typically via freely-accessible web-based platforms. Governments are often hosts, providing access to datasets sourced from various channels on topics ranging from economy and finance to health, education, energy, agriculture, and more. Numerous initiatives have sprung up to advance the cause of open data, such as Open Data for Development (OD4D), the Open Data Charter, the OpenCorporates project, and the G20 Anti-Corruption Open Data Principles and Data Gaps Initiative.[14] Open Data Portals are also increasingly popular in the private sector[15] (e.g., consider Microsoft Research Open Data or Harvard's Dataverse).

The efficiency- and equity-enhancing effects of open data have been documented extensively.[16] In the financial sector, for instance, open data can promote market efficiency and soundness, and also further the publics' understanding of the sector. Open Data Portals hosted by central banks and other financial authorities provide economic and market statistics compiled from surveys, supply-side data (i.e., data submitted by supervised entities as part of their regulatory compliance obligations), and data gathered directly from financial markets (e.g., interest-rate and foreign-exchange data). Firms use these to meet a variety of business needs, such as (i) getting a read of market conditions, (ii) benchmarking their performance, (iii) informing product design, pricing, marketing strategies, and (iv) calibrating risk or valuation models.[17] Without such data, firms must rely on costlier or lower-quality data for their business intelligence, which can put smaller players at a disadvantage. The strong demand for such data has led to the creation of third-party providers who add value to open data and on-sell them in the form data as a service (DaaS).[18]

Open data also improves governance outcomes by empowering consumers and market participants to hold FSPs to account.[19] Accountability, in turn, engenders market discipline, which promotes financial stability.[20] Transparency

13 See: http://opendefinition.org/.

14 See: http://od4d.net/initiatives/, https://opendatacharter.net/, https://opencorporates.com/info/about/ and http://www.g20.utoronto.ca/2015/G20-Anti-Corruption-Open-Data-Principles.pdf.

15 For example, consider Microsoft Research Open Data, Harvard's Dataverse, Facebook Data for Good, Google Public Data, World Bank Open Data, and the United Nations OCHA Humanitarian Data Exchange (HDX)."

16 See: http://odimpact.org/.

17 McKinsey Global Institute (2013), Open data: Unlocking innovation and performance with liquid information.

18 For instance, the real-estate tech company Zillow "built" its company on open real estate data. In addition to Zillow, Google Search's exploitation of the public content on the World Wide Web may also be the ultimate example of this, having built an empire using the World Wide Web Consortium (W3C) standards to "organize the world's information" at large.

19 World Wide Web Foundation, Researching the emerging impacts of open data: ODDC conceptual framework, July 2013.

20 As the BIS notes: "Market discipline imposes strong incentives on banks to conduct their business in a safe, sound and efficient manner, including an incentive to maintain a strong capital base as a cushion against potential future losses arising from risk exposures." Basel Committee on Banking Supervision, Working Paper on Pillar 3– Market Discipline, September 2001.

also incentivizes social responsibility and consumer welfare, as when consumer complaints statistics or service quality indicators draw attention to unmet or poorly-served customer needs, thereby influencing purchasing decisions and encouraging businesses to improve their service.[21] Open Data Portals serve non-commercial ends too. Non-government organizations (NGOs) and academia rely heavily on open data for their research projects. And government entities themselves have begun to use Open Data Portals in lieu of internal data sharing arrangements to gather evidence for policy and decision-making.

Notwithstanding these benefits, Open Data Portals often underdeliver on their promised potential. Data are frequently sparse and scattered across multiple platforms without uniform metadata standards or consistent user interfaces (UI) and user experiences (UX).[22] Much are kept closed, either purposely or due to lack of technical capacity. According to the Global Open Data Index,[23] over 90% of government data was inaccessible in 2016, while the Open Data Barometer finds that fewer than one in five datasets are open, with little improvement in the past 10 years.[24] Furthermore, the most commercially valuable data, such as granular, high-frequency market data (e.g., house prices and sales), are rarely available from Open Data Portals and can only be sourced from third-party providers, often at a considerable cost. Even when granular data are available, convoluted data structures and unwieldy formats make it cumbersome to check their validity (i.e., whether granular data adds up to reported headline aggregates). The plethora of firms specializing in the consolidation and standardization of free public data for a fee testifies to this data management pain point.[25]

In addition to limited breadth, depth, timeliness, and quality of data on Open Data Portals, poor presentation of the data can also undermine ODP's user engagement. A study by Open Data Watch compared web traffic to Open Data Portals with national statistics offices' (NSOs) website showed far lower traffic for the former than the latter, explored Open Data Portals for longer and with greater page depth.[26] At least partly to blame for the low uptake are clunky interfaces and a scattershot approach of publishing reams of raw datasets without due consideration to need or usage. In particular, it presupposes the capability of users to turn data into meaningful and actionable insights. Yet those on the deficit side of the data divide often lack the technological and human resources (e.g., data infrastructure, data scientists, etc.) to properly leverage the data for their ends.

Overcoming these shortcomings is not straightforward. To the extent that they reflect limitations of the data architectures that underlie the platforms- limited storage and computing power, slow turnover times, etc. - creating Big Data capabilities would require large investments into data infrastructure by ODP hosts. For government agencies, budgets for such initiatives tend to be small.

21 For instance, the Competition and Markets Authority (CMA) in Great Britain since 2018 publishes surveys of major banks on service quality. See CMA.

22 There are, of course, exceptions, and open data providers such as CKAN have greatly improved the US data.gov or EU Open Data Portals).

23 See: Global Open Data Index.

24 World Wide Web Foundation, Open Data Barometer: From Promise to Progress, September 2018.

25 Consider, for instance, the case of open data on insurance markets in Latin America. Latinoinsurance provides collects, cleans, and standardizes open financial data from the regional superintendencies of insurance for a subscription fee, in addition to other services.

26 Open Data Watch (December 2018), Measuring Data Use: An Analysis of Data Portal Web Traffic.

Improving user experience and engagement by providing not only raw data but also meaningful insights entails similarly costly investments in interface design and data analytics. Plugging data gaps is further complicated by the narrow scope of government agencies who collect, consolidate, and publish market statistics. Much valuable data is either confidential or out of their remit. This makes the goal of enriching the data offering beyond government data contingent upon the goodwill and commercial interest of data originators. However, these actors may be reluctant to open their data for little or no return, especially if it involves sharing sensitive commercial data with competitors. The free-rider dilemma ("Why share mine if I can use yours?) and collective action problem ("The data is only meaningful if we all [or a sufficiently representative sample] share our data.") work against the objective of Open Data Portals. For these reasons, Open Data Portals are frequently under-stocked and under-utilized.

From Open Data Portals to Open Data Commons

To address these shortcomings of Open Data Portals, the Open Data Commons solution proposed in this report goes several steps beyond simply opening data and forays into establishing a shared data platform for the generation, exchange, and dissemination of insights to serve specific use cases. The notion of a "commons" implies not only openness but also a place for users to share, pool, and exchange datasets. To incentivize such behavior, the platform furnishes its users with some amount of storage space, computing power, and ready-to-use data models and analytics tools at no or nominal cost, in addition to the opportunity to mix-and-match their data with data produced by their peers. This Open Data Commons structure has gained prominence in the scientific community where data volumes are prodigious, research budgets are tight, and the returns to data sharing are large.[27] The success of data commons in this space is partly due to the fact that they target specific use cases, such as genomic sequencing, climate modeling, and earth satellite imaging.

In order to replicate the success of Open Data Commons in scientific communities, and to achieve the vision of an Open Data Commons that bridges the emerging data divide in financial services, three conditions need to be satisfied. The platforms must be able to:

1. Secure a sponsor willing and able to provide, in effect, a public good with nontrivial set-up and maintenance costs of an appropriate data infrastructure.
2. Align with a data architecture that can accommodate Big Data and otherwise address the data shortcomings of Open Data Portals.
3. Create the right incentive structure for data haves to supply data and for data have-nots to make effective use of it.

The solution to meet these conditions proposed here rests on a concept called the DataStack. **The DataStack is a modular, streamlined, end-to-end data architecture that leverages an interoperable data platform and advanced analytics tools to generate meaningful, actionable insights in digestible formats for multiple personas**. Although it is not restricted to any type of organization, a DataStack is ideally suited to governments or government agencies with large

27 Robert L. Grossman, Allison Heath, and Mark Murphy, A Case for Data Commons: Towards Data Science as a Service, January 1 2016.

data resources and needs. Here we focus specifically on financial authorities as the starting point for building an Open Data Commons.

Why start with financial authorities? First, they are the locus of much data that is collected by governments. Central banks, for instance, gather copious amounts of data on the economy, financial sector, and society in order to fulfill their monetary policy and supervisory functions. Furthermore, making data public falls within their mandate. This pool of data can provide the critical mass to launch an Open Data Commons platform.

Second, as argued above, such data is highly coveted by the private sector for commercial reasons, and therefore has a ready user base to "go live" with.

Third, a DataStack solves for multiple use cases in the domain of regulatory and supervisory technology (RegTech and SupTech respectively), and to that extent it is an intrinsically worthwhile undertaking even without the open data component. A DataStack can serve as a unified and rationalized Big Data architecture to service financial authorities such as prudential supervisors, payments oversight and financial inclusion departments within central banks, anti-money laundering authorities, and more.

Fourth, the use cases that are developed for "internal" personas also satisfy external market needs. For instance, the insights extracted from Big Data analytics and AI models for financial inclusion departments can also benefit financial service providers in devising strategies for targeting low-income or excluded populations (see box 1). Analytical dashboards can have both internal and market-facing interfaces with differentiated access rights and user privileges to protect sensitive information.

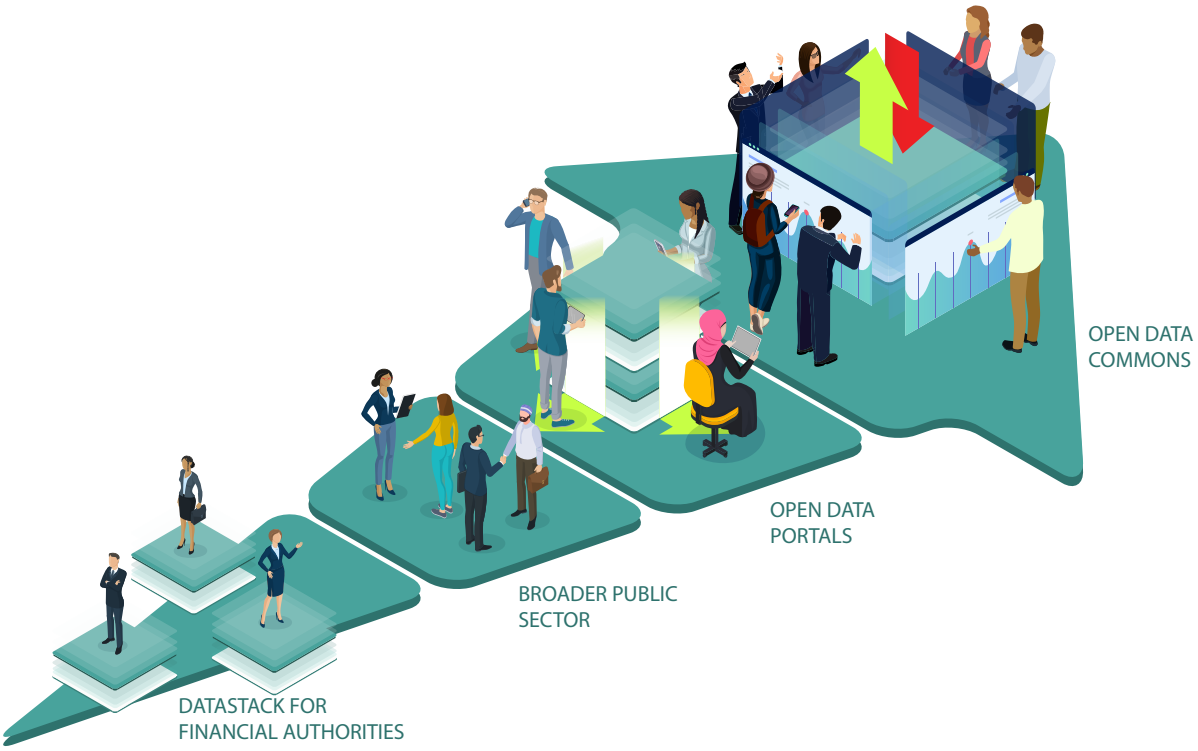
Fifth, the DataStack satisfies market demand for both rich data and meaningful business intelligence, as evidenced by the high price they command in the third-party data marketplace. By crafting experiences around specific use cases - in addition to providing the usual data dumps - user engagement should improve. The value proposition for users to share their data rests on the access to Big Data storage and computing capacity, provided at no or subsidized cost by the government hosts, as well as the opportunity to draw insights (via interactive dashboards and maps, reports, embeddable graphics for mass media. etc.) from a larger pool of data, with the DataStack providing the critical mass of data to start. Use cases also provide a focal point to coordinate data sharing by external users.

Finally, a pay-for-play model similar to existing third-party data marketplaces can also be envisaged as an additional enticement to data contributors.[28]

Once a firm foundation for the DataStack is established within financial authorities, other public-sector personas can be added to the platform. These include line ministries, monetary authorities, credit bureaus, and other government agencies with overlapping data needs and offerings. The DataStack would therefore form a centralized data repository and analytics platform to facilitate data sharing across jurisdictional lines. It would establish a common base of evidence to inform and coordinate public policies across the public sector. Bringing in other government agencies onto an interoperable data platform also forces the creation of common metadata standards that can then be extended to users outside of the public sector.

28 For example, Google Cloud's Commercial Datasets: https://cloud.google.com/commercial-datasets

FIGURE 2: THE DATA SHARING JOURNEY BASED ON THE DATASTACK BLUEPRINT



BOX 1: EXAMPLE OF OPEN DATA COMMONS USE CASE - MONITORING FINANCIAL SERVICES AGENTS

Financial service providers managing networks of banking, mobile money, or insurance agents need to monitor agent location to inform operational decision making. Such monitoring can: (i) guide the rollout and distribution of new agents; (ii) suggest product and service distribution based on local economic conditions; (iii) help plan the most efficient routes for runners to support agents; (iv) and improve the efficiency of liquidity management. Furthermore, real-time monitoring can help ensure that agents remain within their designated areas, rather than cluster around the most lucrative locations.

Advanced GIS (geospatial information system) tools can confer a significant competitive advantage on providers with the requisite infrastructure and tools to apply them. Often providers invest significant resources into building futuristic command centers with live maps and dashboards of their agent networks. Without effective agent monitoring capabilities, FSPs may lack timely intelligence about their agents' positions, potentially leading to an inefficient distribution of agents and a misallocation of resources. This asymmetry can hurt competition and degrade the quality of service, especially where an incumbent has preponderant market power.

Opening market-wide data about agent locations can benefit financial authorities, providers, and consumers alike. As explained above, supervisors can use live agent monitoring to enforce location requirements or aid financial inclusion efforts by identifying coverage gaps. Smaller players or startups gain access to powerful analytics that can improve operational decision-making and contribute to a more level playing field. Market leaders and incumbents will have richer data with which to optimize their distribution strategies. In a non-exclusive market (i.e., multiple tills per agent), players will be able to identify agents by their affiliations and vie for their business. Greater competition and improved delivery should translate into lower prices and better-quality services for consumers.

The data requirements for such a solution must include some location marker that can be updated periodically, ideally in real-time. Often location is determined simply by the agents' self-reported addresses of registration, which cannot be independently or dynamically verified, and are less precise than geolocation. The latter uses geographic coordinates and GIS for live positioning. Some providers are equipping agents with smartphone apps that send live positions to the network manager.

Where agent locations are part of regular regulatory filings and can be shared openly, a market-wide view of the agent network can be generated from supply-side data. Where not, providers need to be incentivized to share their agent data. An Open Data Commons use case would entail creating a portal for market participants to upload their agent location data onto the DataStack platform, plot them on maps, combine them with their peers' data, and overlay them on contextual information such as demand-side surveys and utility maps. Providers may be reluctant to share sensitive operational data with competitors, yet they could be enticed by the prospect of gaining a holistic market view plus access to analytical tools.[29] The DataStack would save them the considerable time and resources required for building a GIS tool for agent mapping and training staff on GIS analysis, particularly more complex spatial machine-learning techniques such as segmentation and clustering.[30] And they would gain access to contextual and market data from which to extract insights.

29 One possibility is to give providers the option to reveal their locations in exchange for unhiding the others'.

30 See for example: Rohit Singh (2018), Integrating Machine Learning and Deep Learning with ArcGIS.

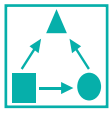
The DataStack: A blueprint for the Open Data Commons

The DataStack establishes the foundation upon which to build an Open Data Commons. Here we unpack the concept into its seven core features:

Personas are archetypal end-users of the DataStack. The base personas for our initial examination are financial authorities and related government agencies. These personas can encompass different functional units of the same institution, such as prudential supervision and payment system oversight within the central bank, as well as institutionally separate entities such as conduct and competition authorities and financial intelligence units (FIUs). Outside of this core, the personas can refer to any entity, public or private, that can benefit from data-driven public policy and decision-making. Open Data Commons personas might be financial services providers, non-governmental organizations, donors, researchers, and ordinary citizens and companies. The personas determine the use cases, which drive the definitions for the data and functional requirements.

Data architecture comprises a data platform, data models, and front-end interfaces and analytics tools that collect, process, store, and render data from multiple sources and in varied formats. The architecture is optimized for performance, cost, and efficiency by selecting components that allow hosts with tight budgets to still take advantage of advances in Big Data and AI. Cost and efficiency savings are achieved by: (i) eliminating duplication of processes; (ii) automating manual tasks to reduce costly human errors; (iil) leveraging open-source technology to avoid licensing costs; (iv) building users' capacity to operate and maintain their tech in-house, thereby lowering maintenance costs; and (v) leveraging analytical tools to squeeze the most value out of data and existing resources.

The DataStack is **modular** in that it is pieced together using technology components that are best suited to meet the particular technical and functional requirements of users. A DataStack is not an off-the-shelf "software stack" or "software as a service" (SaaS) product, but rather a menu of best-in-class technologies that address the idiosyncratic institutional and budget constraints of the host organization.



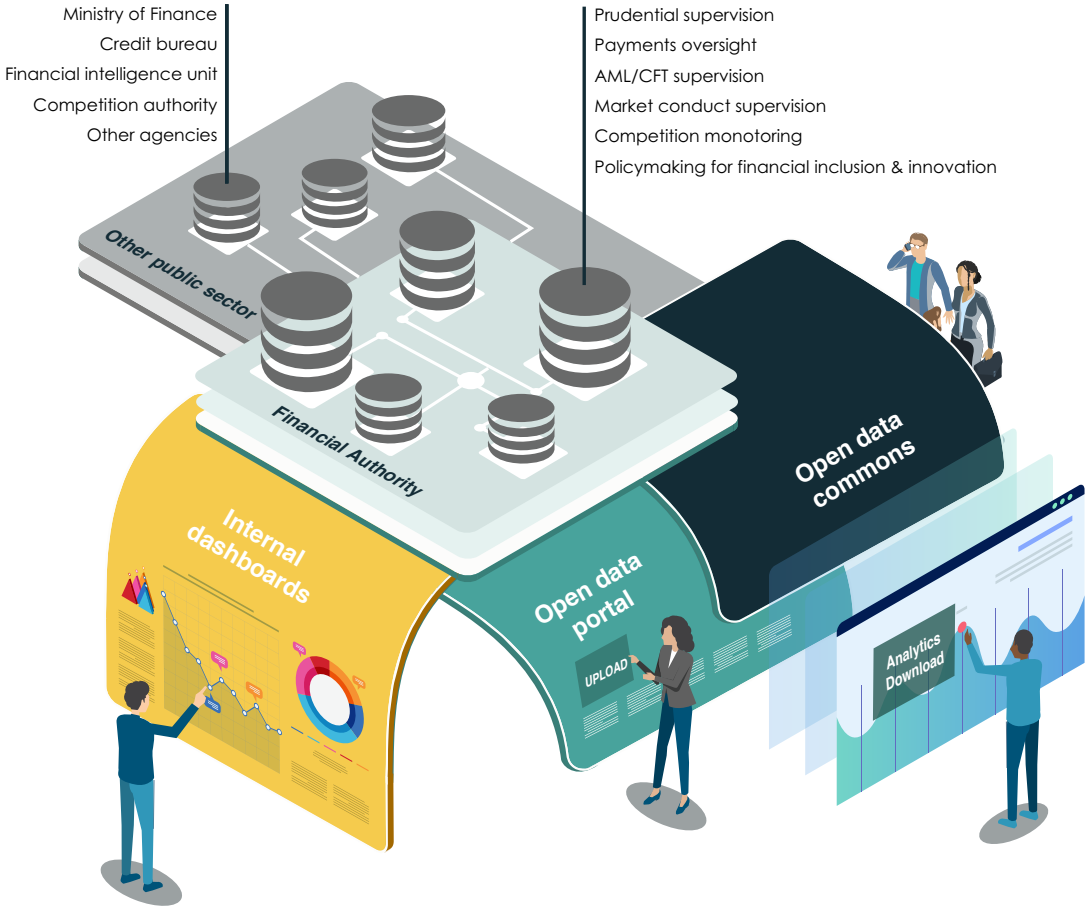
To be **interoperable** across databases and to accept data sets from external contributors, common standards for data types and file formats in order for the platform are required. These standards should include taxonomies and ontologies that index and parse incoming data for easy search and recovery. The standards can be determined when creating the core DataStack or when it is connected with other government agencies, but ideally there should also be scope for community participation in standard setting in order to entice data contributions (i.e., avoid ensuring standards are achievable by potential contributors).

Insights are meaningful and actionable descriptions and explanations of phenomena, in the primary instance financial and economic ones. For public sector personas, they serve as the basis for crafting regulations and supervisory and public policy actions, thereby helping decision-makers fulfill their duties and meet their objectives. For private sector personas, they help to inform strategic decisions around product design, marketing and customer engagement, the formation of startups, and other business processes. Insights are generated using analytical tools including Big Data and AI applications such as natural language processing (NLP), anomaly detection, robotic process automation (RPA), sentiment analysis, predictive analytics and machine-learning.

Digestible formats include interactive dashboards, maps, and reports that incorporate best practices in UI and UX design so that critical information is communicated quickly and effectively. To this end, DataStack architects should leverage the large body of work from consumer and business insights that has been built for the private sector.

Security is a cornerstone of the DataStack. The extent to which hosts can guarantee the confidentiality and protection of intellectual property will determine the degree of openness that they can and should allow. Strong user access and identity management features as well as best-in-class security protocols for data in transit and at rest are therefore indispensable. To safeguard commercially-sensitive data, strict and secure access control needs to be built in, granting viewing rights only to those datasets that data contributors wish to make public. Hosts may require prior registration and authorization of external users to control who has access, in particular where they are able to contribute data and therefore potentially introduce security risks.

FIGURE 3: INTRODUCING THE DATASTACK



The DataStack: A blueprint for financial authorities

Financial authorities need a new data architecture

The digitization[31] and datafication[32] of vast swathes of the economy are increasing the size and complexity of the policy domains that regulators, supervisors, and other government agencies administer. Their purviews have expanded to encompass new mandates and policy objectives, from protecting consumers to promoting sustainable finance, thereby creating new data needs while adding to an already heavy data management workload. Best practices in risk-based supervision, proportional regulation, and evidence-based policymaking, which rely on data-intensive models and methodologies, are also fueling a demand for data. While these models can be costly or complicated to integrate into existing data architectures, a regulatory and policy environment marked by heightened uncertainty and complexity makes accurate, granular, timely, and trustworthy data ever more necessary.[33]

31 Digitization is the process of converting information into a digital (i.e. computer-readable) format, in which the information is organized into bits. https://whatis.techtarget.com/definition/digitization . https://www.collinsdictionary.com/dictionary/english/digitize

32 Datafication is a technological trend turning many aspects of our life into data[1] which is subsequently transferred into information realised as a new form of value. Cukier, Kenneth; Mayer-Schoenberger, Viktor (2013). "The Rise of Big Data". Foreign Affairs (May/June): 28–40.

33 See, for instance: International Monetary Fund (2018), Global Financial Stability Report: A Decade after the Global Financial Crisis: Are We Safer?.

Many data systems that underpin regulatory, supervisory, and public policy frameworks are poorly suited for the abundant supply of data. Indeed, data overload renders many data management strategies and systems obsolete or even counterproductive. Numerous diagnostic studies of financial authorities and government agencies performed by BFA Global well as surveys conducted by our RegTech for Regulators Accelerator (R^2A) reveal how inefficient, outdated, and redundant data management systems and processes create debilitating pain points for their users.[34] Common examples include:

The pervasive use of Excel spreadsheets in the data management workflow, from preparing reporting templates, to storing historical data and generating reports. While Excel works well for small datasets, its limited memory and computing power circumscribe the kinds of data analysis that can be performed.

Email remains the dominant method of transmitting data and sharing files, even though it too imposes strict size limitations and is inherently risky in terms of data integrity and general security.

Manual and paper-based processes are also ubiquitous, from physical filing of license applications to validation via "eyeballing". Manual processes are tedious, time-consuming (costly), and error prone.

Disjointed and non-interoperable data systems slow or staunch the flow of data within an organization, creating bottlenecks and curtailing efficiencies of scope.

Together these pain points can lead to regulatory bottlenecks, supervisory blind spots, delayed reaction times, superficial diagnoses and blunt interventions. Financial innovation may be discouraged by lengthy and cumbersome licensing and reporting requirements. Lapses in oversight means systemic risks can lurk undetected until they escalate into systemic shocks. A lack of context or nuance may create disproportionate or misguided policies and regulation. At a minimum, these shortcomings undermine efforts to improve market competition and promote financial innovation and inclusion. At worst, they can engender or exacerbate financial crises.

On the other hand, upgrading and modernizing all or critical parts of the data architecture can unlock significant efficiency gains and improve regulatory, supervisory, and public policy outcomes. New technologies to that effect, known respectively as SupTech, and GovTech, promise to: improve data management and analytics, inform business intelligence, guide decision-making, and direct policy actions. Big Data tools such as application programming interfaces (APIs), RPA, data warehouses, and "smart" dashboards can dramatically increase the size, speed, variety, and integrity of data. Powerful analytics tools such as ML, optical character recognition (OCR), NLP, and other AI methods have been increasingly commoditized and can be responsibly embedded to mine existing and untapped quarries of data, such as responsibly-collected web content or mobile phone application use, for meaningful and actionable insights. The DataStack combines cutting edge, cost effective, and appropriate GovTech and SupTech solutions into a coherent and streamlined technology stack.

34 See: BFA Global RegTech for Regulators Accelerator (R^2A) (2019), The State of RegTech: The Rising Demand for "Superpowers".

Adaptability by design

Although the DataStack is designed to maximize the viable portability of features, each incarnation of the DataStack is nonetheless unique in its nuances and thus requires some amount of custom development work to tailor it to the users and data sources. A wholesale reengineering of a given authority's data architecture is rarely feasible. To start, different jurisdictions have different degrees of freedom in changing their data architectures. Replacing or upgrading systems is generally expensive, especially when the systems lock users into costly service agreements, require heavy investments for on-boarding and training, or form part of exclusive software stacks. Furthermore, legacy agreements and infrastructure can create path dependencies that are difficult to reverse. Authorities may be reluctant to write off sunk costs or retrain personnel. Institutional inertia and bureaucratic politics typically stymie efforts to rationalize data management processes. Finally, departments might be overly protective of their data when concerns about job security come into play, or limited by existing regulatory requirements. In all, there tend to be many reasons that make change difficult.

Overcoming these obstacles to change requires proper coordination, sequencing, and pacing. It often makes sense to implement bite-sized pieces that add up to a full-scale DataStack over time. One approach employed by BFA Global and R²A starts with a holistic diagnosis of a given data architecture to identify specific points of inefficiency, redundancy, and vulnerability that might be addressed through one or more SupTech, and GovTech applications.[35] Individual use cases are then evaluated and ranked in terms of feasibility, impact, and value/priority. Development then proceeds in sequence or concurrently, ensuring that each alteration or addition fits into a coherent structure and allows for future adjustments.

Taking a diagnostic approach has the advantage of being grounded on the observed and felt experiences of the current and prospective users, in this case, the staff of authorities and agencies directly involved with data management. This way, users are directly involved in crafting the solutions and tailoring them to their particular needs and circumstances, as opposed to deploying off-the-shelf solutions that often are not fit for purpose. It also provides a roadmap for constructing the DataStack in an incremental yet deliberate way.

Building an "open" architecture that allows access to data can contribute to sustainability. Open access can generate network effects: The more people use a shared platform, the more useful it becomes. As more data sharing occurs, users become accustomed to and dependent on the ready availability of data in their work. While data sharing within the DataStack is primarily targeted at functional units within an institution, the same is true for opening data to the public. Granting access to certain data and sharing meaningful and actionable insights - after stripping out confidential or sensitive information - can make external stakeholders more forthcoming with their data, and possibly more amenable to comply with regulatory reporting requirements.

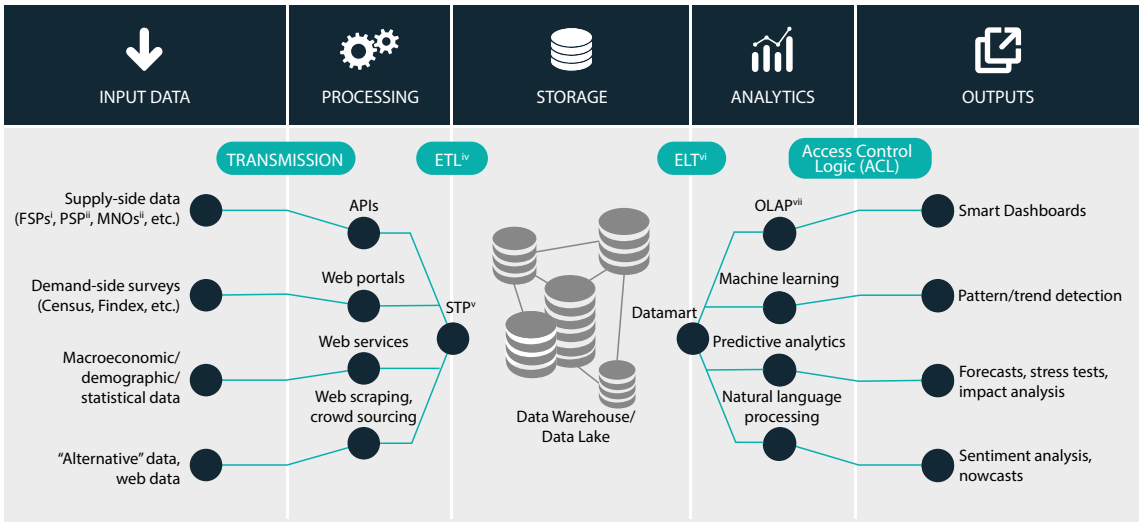
Adaptability, or the DataStack's ability to evolve over time and integrate seamlessly with different systems and applications, also drives sustainability. This agility can

35 See: Simone di Castri, Matt Grasser, and Arend Kulenkampff (2018b), The RegTech for Regulators Accelerator (R²A) Process: Giving Financial Authorities Superpowers, BFA Global.

be achieved by relying on open-source software and open-standard protocols, and by allowing exports to spreadsheets and to plug third-party software. While spreadsheets are the root cause of many pain points, they remain deeply popular and effective as a basic analysis tool. For loyal spreadsheet users, the DataStack's value proposition lies in the easily accessible integrated database. Keeping the DataStack open and nimble is essential to avoid path dependencies and create a strong foundation for an Open Data Commons.

Adaptability is also crucial given the structural transformations currently reshaping the economy and financial sector, which create new and more pressing data management needs. These are: (i) the digitization and datafication of large swathes of the economy; (ii) the meteoric rise of fintechs and TechFins; (iii) strong growth in financial services, digital or otherwise; (iv) a push towards regulatory modernization and global harmonization of standards; and (v) a shift in risk environment toward heightened volatility, uncertainty, complexity, and ambiguity. Together these trends imply a profusion of new data that regulators, supervisors, and policymakers will need to factor into their decision-making - and that data architects will need to incorporate into their designs. Existing technology may not be suitable or capable of handling such volumes, velocities, and varieties of data, and even when technologies are emerging to meet these needs - such as SupTech, and GovTech - existing data architectures struggle to accommodate their requirements. These technologies may utilize untapped sources of data such as TechFins, unstructured types such as web content, novel formats such as images and tweets, as well as new channels such as mobile crowdsourcing or crowdsensing (more on these below). Most data architectures and technologies are not capable of leveraging these sources.

FIGURE 4: Stylized DataStack blueprint



i. FSPs = Financial service providers
ii. PSPs = payment service providers
iii. MNOs = mobile network operators
iv. ETL = "extract, transform, and load"
v. STP = straight through processing
vi. ELT = "extract, load, transform"
vii. OLAP = on-line analytical processing

BOX 2: CASE STUDY - NIGERIA FINANCIAL SERVICES (NFS) MAPS

BFA Global piloted the DataStack in Nigeria from 2017 to 2019. The Central Bank of Nigeria (CBN) and the Nigeria Inter-Bank Settlement System (NIBSS), the country's payment switch, had engaged the Bill & Melinda Gates Foundation (BMGF) and BFA Global to help redesign their data architecture to improve payments oversight and financial inclusion. The solution, Nigerian Financial Services Maps (NFS Maps), created a proof of concept (POC) for the DataStack that we can share with other countries and personas. The project also validated the process by which a DataStack might be implemented in different contexts.

The NFS Maps project grew out of the BMGF's Geospatial Tool for Financial Inclusion Analysis, which used GIS to improve the measurement and tracking of financial access points.[36] The DataStack transformed this mapping system by adding additional analytical layers and a more robust backend. SupTech solutions developed under the BFA Global R^2A initiative also helped to flesh out the NFS Maps personas to incorporate each use case as a leg of the DataStack. The prototypes also validated the modular approach to structuring data architectures, and illustrated the benefits of using SupTech to solve data-intensive regulatory and supervisory challenges.[37]

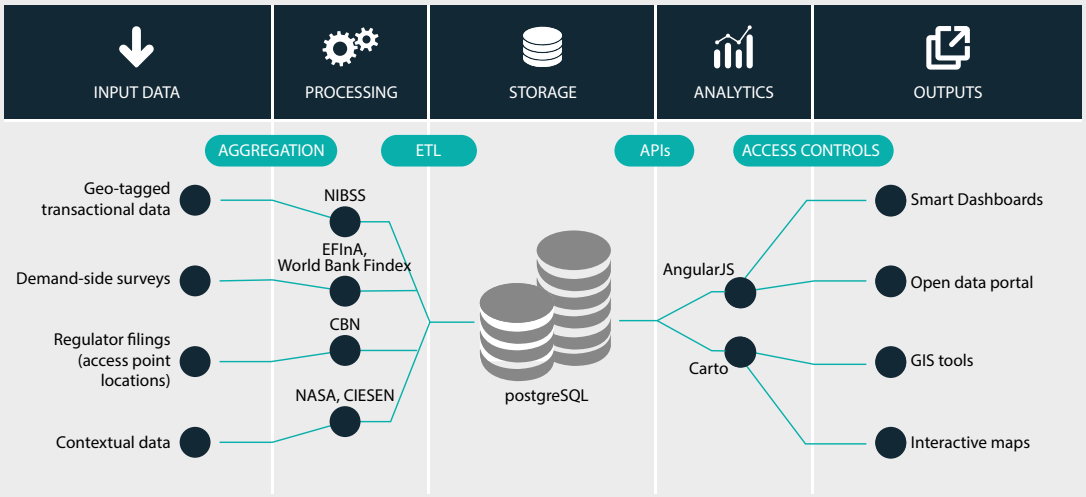
BMGF, BFA Global, CBN, and NIBSS experts met in late 2016 and early 2017 to brainstorm and prioritize possible regulatory use cases for geospatial and payments data. The group evaluated the current status of payments infrastructure, including data flow, staff skills, and other resources, and based infrastructure, including data flow, staff skills, and other resources, and based on this assessment, explored applications for payments system oversight, AML supervision, policy impact analysis, government-to-person (G2P) schemes, and others. At the conclusion of the workshop, the group crafted a shared vision for a data-driven dashboard solution to be led by NIBSS, and synthesized a work plan for building a prototype.

The first incarnation of the DataStack product was delivered in April 2019 and focused on two use cases - payment system oversight and financial inclusion. It also incorporated the wishes of several other departments, including consumer protection, and financial policy and regulation. The data architecture of the product is sketched out in Figure 5.

36 See Brian Loeb and Abed Mutemi (2016), Building sustainable geospatial data resources for financial inclusion, Insight2Impact paper. And http://www.fspmaps.com/#/Nigeria/finance/map@9.31,7.93,z6,dark.

37 Simone di Castri, Matt Grasser, and Arend Kulenkampff (2018a), cit.

FIGURE 5: ARCHITECTURE OF THE NIGERIAN DATASTACK



NIBSS provided the platform with granular transactional data, and the BFA Global team synthesized it with complementary and contextual data supplied by CBN via APIs established for that purpose. All production data was hosted by NIBSS servers on their premises, and was stored in a SQL data warehouse. A third-party GIS vendor, Carto, was used to render the maps.[34]

The payment system dashboard depicted key performance indicators for Nigeria's payments system in the form of risk dials, time series charts of transaction trends and KPIs, and a chart of payment access points. The map plotted access points and enabled deep dives into individual merchants and agents, both bank and mobile, enabling the supervisor to quickly view transaction histories, consumer complaints records, and incidences of fraud, theft, and robbery for each payment point. Other features included real-time tracking of KPIs, such as failure rates of point-of-sale (POS) devices, and NIBSS instant payments (NIP).[38]

For the financial inclusion use case, the NFS Maps augment the geo-mapping demand-side financial access survey with supply-side data collected by CBN and NIBSS, in addition to adding new contextual layers such as crime and political risk maps. This provided a rich and dynamic overview of the Nigerian financial ecosystem and enables CBN's Financial Inclusion Secretariat to more effectively pursue its FI strategy, which includes deploying 500,000 mobile money/bank agents by 2020.[39] Furthermore, including complaints data and crime statistics helped the central bank enhance the consumer protection mandate created in the latest National Financial Inclusion Strategy.

An open data portal provided access to both maps and dashboards for public browsing. Whereas public users have unrestricted views of the ecosystem maps, drill-down features and historical data export were disabled so as to protect sensitive commercial data. In addition, a separate, access-controlled, market-facing portal allowed FSPs to compare their geographic profiles against the market (albeit with the names of their competitors obscured).

38 See: Carto.

39 See: NIBSS live data.

Conclusion

The DataStack corrects the emerging "data divide" that threatens to skew the development paths of financial sectors in favor of players with the means to manage and mine the data being generated by digital financial services. An Open Data Commons that rests on a core DataStack centered around financial authorities and government agencies would provide users with access to data, computing power, storage facilities, and analytical tools that together surface meaningful and actionable insights for decision-making.

The digital divide that has emerged alongside the digital and mobile transformation of financial services is starting to close. In its place, a new *data* divide is opening. Those with access to rich mines of data and who are able to effectively use them to surface meaningful and actionable insights for decision-making are at a distinct advantage over their competitors and counterparts in government who lack access and analytics resources.

Financial authorities figure prominently on the deficit side of the data divide. Although they are the recipients of substantial amounts of supply-side data via regulatory compliance reporting, they often lack the means to manage and mine them effectively. Data from digital platforms, products, and providers are proliferating at a dizzying pace, frequently overwhelming the ability of regulators and supervisors to capture relevant information and distinguish signal from noise. Meanwhile, these authorities are also assuming new mandates, such as the promotion of financial inclusion and innovation, which brings additional data needs. There remain large mismatches in the demand and supply of data for regulatory and supervisory purposes. Against this backdrop, the failure of financial and monetary authorities to adapt to the emerging reality of data abundance risks misallocating regulatory resources and lapses in financial supervision, potentially undermining financial stability and integrity.

Addressing these challenges and risks requires a new approach to building and structuring regulatory/supervisory data architectures. Existing data architectures have multiple weaknesses and vulnerabilities in reporting and analysis. Inefficient manual processes cause undue delays. Size and processing limitations create bottlenecks. The low granularity of data, limited storage space, and insufficient processing power curtail the application of analytics tools, especially data-intensive Big Data/AI applications. Many off-the-shelf solutions, such as a PaaS or SaaS, lock users into costly service agreements and rigid software stacks that do not adapt to conditions on the ground, especially in lower-income jurisdictions. Nor do they build the capacity of financial and monetary authorities to maintain, service, and adapt their architectures as needs and circumstances evolve, in particular, to incorporate a widening array of Big Data and AI tools.

The DataStack addresses these risks and challenges, and alleviates the pain points of existing data architectures. It offers a modular, interoperable approach that allows various personas to surface insights efficiently and securely in digestible formats. It is technology-agnostic in that it seeks to assemble the most suitable, cost-effective, and innovative parts for every layer and application of the stack. To help financial and monetary authorities harness data profusion, DataStack favors solutions that are capable of handling large volumes of data efficiently

40 See: CBN, National Financial Inclusion Strategy (Revised), October 2018.

and at a high velocity and bandwidth. It facilitates data interoperability to fix coordination failures in the data economy by creating a platform that can integrate different types of data from various sources, and by providing a means to share and disseminate that data for wider use. Better data management and the application of cutting-edge SupTech solutions, in turn, will generate insights that can inform and guide personas of all types, be they policymakers at financial and monetary authorities or private sector analysts.

In fact, by reducing the data gaps, the DataStack creates value across the whole financial ecosystem with the public sector as the catalyst.

BFA GLOBAL

R²A

REGTECH FOR REGULATORS ACCELERATOR

Contact

R2A@bfaglobal.com

www.bfaglobal.com/project/R2A

@BFAGlobal | @R2Accelerator

www.linkedin.com/company/bfaglobal