Data Sharing - Meeting the Challenge and Maximising the Opportunity

Dr Ilesh Dattani CEng CISA

Assentian Limited

Done well, data sharing brings numerous benefits. It creates an ecosystem of organisations that can collaborate together to achieve common goals. Data sharing also plays a critical role in building knowledge for communal benefit. Governments at the federal, state, county, and municipal levels have become increasingly interested in the treasure troves of local data amassed by the sharing economy.

Data drives scientific and technological breakthroughs, underpins policymaking, and powers the global economy. Clinicians use data to identify the best treatments for their patients, farmers use data to predict and improve farm yields, researchers use data to generate new knowledge about natural and social phenomena, and public servants use data to create evidence-based policies.

Artificial Intelligence (AI) and other emerging analytics techniques are amplifying the power of data, making it easier to discover new patterns and insights ranging from better prediction models to understand and mitigate the impacts of climate change to new methods for detecting financial crime.

Although data enables science, innovation, and insights, balancing the benefits of these data-derived insights with the imperatives of privacy, security, and other values is a longstanding challenge. For example, when developing new treatment options, medical researchers may benefit from broad access to electronic health records. However, those records may contain personal health information related to individual patients, compromising the privacy and safety of those patients as well as rights under health privacy laws and regulations on the protection of human subjects. Similarly, when researchers access authorized data without safeguards on how they access the data, privacy-sensitive information such as their location or the specific type of information they are accessing may be revealed. In many domains, collaborations that could improve Al model training and accelerate progress must be balanced with ethical and legal privacy concerns and intellectual property protection concerns.

Data sharing challenges

Many data owners have difficulties in agreeing to share data. They may see data sharing as too risky because of privacy and security challenges. They may be concerned about losing control of "their data" and it being used inappropriately or disclosed to other parties. Understanding the challenges can help data owners to find solutions that overcome them.

Data sharing challenges in the private and public sectors

While data sharing is well-established in research institutions where it is a vital part of verifying results and conclusions, many data owners in private and public sector organisations feel challenged when sharing data for the first time.

Some of these challenges include:

- Data sensitivity
- Organising data in a presentable, sharable, and useful way
- Accidental disclosure of confidential or personal information
- Deliberate disclosure of data by a bad actor
- Violation of intellectual property rights and other interests
- Loss of control over data
- Costs of sharing data
- Misuse of data for unauthorised purposes

Why you need to address data sharing challenges

Data sharing has several benefits for progressive organisations and is considered critical for projects that need collaboration between stakeholders. Data sharing also contributes to the future success of a business by making information available and accessible. This data can then be reused and reinterpreted in ways that address new areas of understanding, such as in developing better government policy or enhancing consumer services.

By taking the right precautions and insisting on proper data management and organisation, data sharing can become a smooth and streamlined exercise.

The ASEAN Context

The ASEAN countries as a regional group have created the ASEAN Framework on Digital Data Governance.

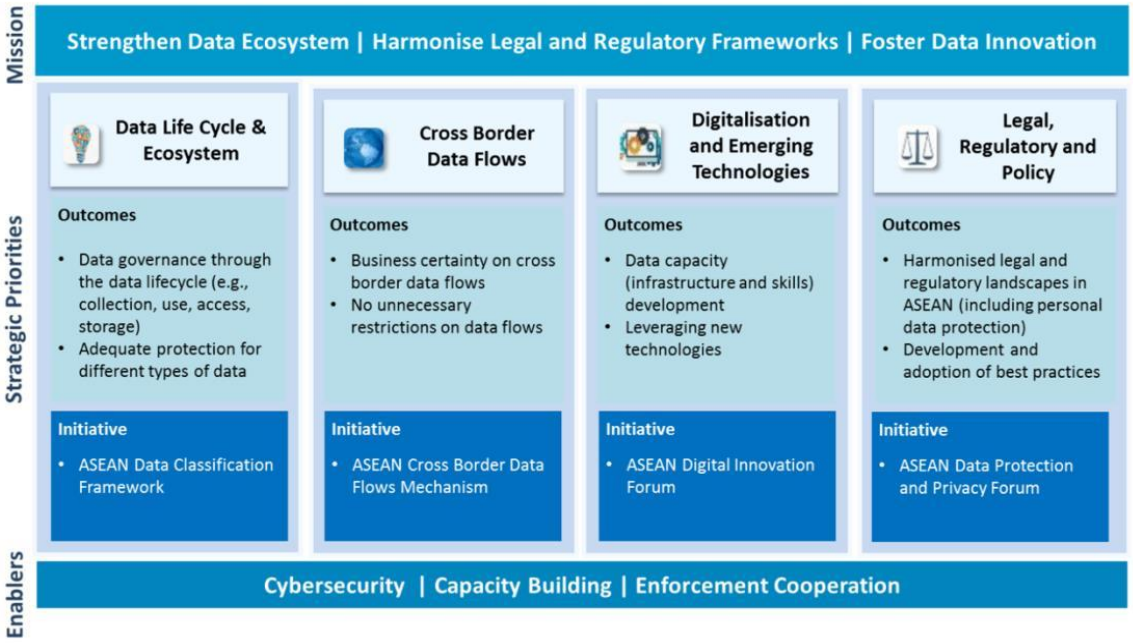

Figure 1: Summary of the ASEAN Framework on Digital Data Governance

If we look at each of these 4 pillars as shown in the figure above, we see considerable overlap between the strategic priorities and the three overarching missions. Digitalisation and Emerging technologies are a facilitator of data sharing and can deliver the outcomes desired within the data life cycle & ecosystem and cross border data flows priorities - as such digitalisation and emerging technologies is a cross cutting enabler. Legal, Regulatory and Policy initiatives can often become barriers to the very disruption, innovation and enablement being sought - as such it requires cooperation across stakeholders and technology is again a vital enabler for meeting the challenges that come from regulatory controls.

Data Lifecycle and Ecosystem

Data Integrity and Trustworthiness: Knowing where your data comes from, who has seen it, what has been done with, how has it changed and where it has been is crucial to maintaining the integrity of the data and any outputs that may be derived from it as well. Without the required level of transparency, trust is eroded in the data and for those individuals and/or organisations looking to share data the opportunity and/or benefits are potentially outweighed by the risks they foresee as a result.

Data Use and Access Control: It is imperative that unauthorised person(s) and/or organisations are not able to access and utilise data that an individual and/or organisation may have consented to share.

Data Security: Safeguards for Data at Rest, in-transit and when it is in operational use are all required and approaches should be guided by what classification the data has and how critical it is to the owners.

Cross Border Data Flows

Data Innovation is dependent on large volumes of data and is increasingly a requirement for interoperability making people's lives easier as they become increasingly mobile. Data Sharing within the country and between countries fosters a more vibrant data ecosystem and facilitates regional and international collaboration on innovation. Appropriate safeguards need to be in place so that individual rights are maintained and this links directly to the regulatory and policy making pillar where appropriate safeguards on Data Privacy and Data Protection need to be defined. Any such regulatory framework needs to ensure equivalence between the ASEAN Member countries and a common regulatory framework generally eases this as is the case with the European Union General Data Protection Regulation (GDPR).

Digitalisation and Emerging Technologies

Technology is an enabler for Data Sharing and emerging innovations provide mechanisms by which Data Sharing can be done at scale whilst providing all the necessary safeguards as envisaged within the ASEAN Framework on Digital Data Governance.

The key enabling technologies are:

Data Lineage

Data lineage is the process of tracking the flow of data over time, providing a clear understanding of where the data originated, how it has changed, and its ultimate destination within the data pipeline.

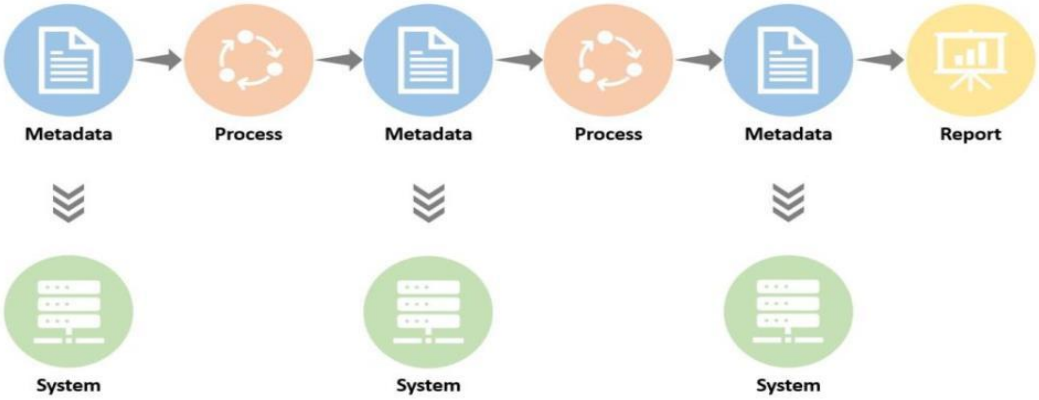

Figure 2: The Data Lineage Process across the data lifecycle

The notion of Data Lineage is not new however emerging technologies like Machine Learning have accelerated its ability to provide the level of granular insight required to maintain integrity and trust in data which is a key enabler for data sharing. The above figure shows how Data lineage can be seen as a type of metadata that traces relationships between upstream and downstream dependencies in your data pipelines. Lineage is all about mapping: where your data comes from, how it changes as it moves throughout your pipelines, and where it's surfaced to your end consumers. In addition to providing the capability to maintain integrity and trust in data it also can be a core component of the data innovation workflow.

- Understand how changes to specific assets will impact downstream dependencies, so they don't have to work blindly and risk unwelcome surprises for unknown stakeholders.
- Troubleshoot the root cause of data issues faster when they do occur, by making it easy to see at-a-glance what upstream errors may have caused a report to break.
- Communicate the impact of broken data to consumers who rely on downstream reports and tables - proactively keeping them in the loop when data may be inaccurate and notifying them when any issues have been resolved.

With the right metadata for a given data asset included in the lineage itself, you can get the answers you need to make informed decisions:

- Who owns this data asset?
- Where does this asset live?

- What data does it contain?
- Is it relevant and important to stakeholders?
- Who is relying on this asset when I'm making a change to it?

When this kind of contextual information about how data assets are used within your business is surfaced and searchable through robust data lineage, incident management becomes easier. You can resolve data downtime faster, and communicate the status of impacted data assets to the relevant stakeholders in your organization.

Distributed Ledger Technologies

Distributed Ledger Technology (DLT) has been gaining traction in recent years as a promising tool to enhance data security and integrity for data management and sharing. DLT is a peer-to-peer network of computers, or "nodes", that records and verifies transactions on a decentralized database. This ledger system is managed autonomously, meaning that it requires no central authority to govern or authenticate transactions.

DLT's decentralized model offers numerous security benefits for data management and sharing. Transactions are recorded and synchronized across a network of nodes, making them tamper-proof and immutable. This ensures the security and integrity of the data, as any malicious attempts to alter the data would be immediately detected and rejected by the other nodes in the network. Furthermore, DLT offers improved privacy and confidentiality of data. Data is stored in a secure, encrypted format and is accessible only to authorized users. This ensures that sensitive patient data is not exposed to unauthorized individuals or organizations.

In addition, DLT provides a secure platform for data sharing across multiple stakeholders. By leveraging smart contracts, data providers can securely and efficiently share data between them with minimal risk of data breach or manipulation. Overall, DLT is a powerful tool to enhance data security and integrity for data management and sharing. As the technology continues to evolve, it is likely to become an increasingly important part of industry's data security landscape.

Healthcare provides a strong example of the vast public benefits that can emerge from Data Sharing and how DLT can bring that about. Societies' wellbeing and outcomes could be influenced as data sharing could aid preventative care and medicine and could as a result improve life outcomes, improve children's learning capacity in socially deprived areas and deliver long-term productivity gains to economies. Data is a commodity and patients, being an intrinsic part of the value chain, could be able to choose other charitable causes and community projects where 5% of the revenues to the healthcare provider can be invested. The solution could capture, aggregate and analyse data to better highlight drivers of regional inequalities in health and social care outcomes, empowering policymakers to more efficiently address these regional disparities.

DLT and encryption could drive innovation in preventative care and community-based healthcare models. This technology is the missing link that could enable a synergy between rising trends in healthcare that could be vital in helping the world improve the health of communities. Some of the foreseen trends that could enable this innovation are as follows:

- Innovative healthcare models: The rise of experiments in healthcare such as Accountable Care Organisations (ACO) and community-based healthcare indicate an increased focus on preventative vs. reactionary healthcare
- New technology and data: The increased utility and prevalence of Electronic Health Records (EHR) and mobile health applications (for example Apple Health) foreshadow a future where individual and aggregate population data has greater utility
- DLT and cryptography: A new method pseudonymously collecting, storing, protecting and sharing health data may be possible without violating the requirements of the Health Insurance Portability and Accountability Act (HIPAA) that have traditionally limited the utility of health data and delayed its use in real-time

Privacy Preserving Data Sharing and Analytics (PPDSA)

Privacy-preserving data sharing and analytics (PPDSA) solutions include technical and sociotechnical approaches that employ certain types of privacy-enhancing technologies (PETS) to generate value from and enable analytics on data while protecting privacy and security. Some PPDSA approaches allow users (e.g., researchers and physicians) to gain insights from sensitive data without exposing the original data itself or allow them to access shared data without being tracked or identified. Other PPDSA approaches enable data sharing by obscuring personal data or making synthetic reflections of the original data that preserve the properties of interest in the data while protecting individual privacy.

PPDSA technologies can unlock new forms of collaboration and new norms in the responsible use of personal data. By enabling the use of more comprehensive and diverse datasets, PPDSA technologies can help the global community tackle shared challenges and drive solutions in areas such as healthcare, climate change, financial crime, human trafficking, and pandemic response and achieve more equitable outcomes for underserved, marginalized, and vulnerable populations. PETs are essential for enabling data sharing and analytics in a privacy-preserving manner. The table below provides an indication of the main technical approaches that can now be used to facilitate data sharing whilst preserving privacy.

Technique	Description	Value	Limitations
K-anonymity	Transforms a given set of *k* records in such a way that in the published version, each individual is indistinguishable from the others	Reduces the risk of re-identification	Vulnerable to reidentification attack if additional public information is available
Differential Privacy	Adds noise to the original data in such a way that an adversary cannot tell whether any individual's data was or was not included in the original dataset	Provides formal guarantee of privacy by reducing the likelihood of data reconstruction or linkage attacks	Limited to simpler data types; challenge in managing tradeoff between privacy, accuracy, or utility of data
Synthetic Data	Information that is artificially manufactured as an alternative to real-world data	Preserves the overall properties or characteristics of the original dataset	May still disclose privacy-sensitive information contained in the original dataset; difficult to mirror real-world data
Secure Multiparty Computation	Allows multiple parties to jointly perform an agreed computation over their private data, while allowing each party to learn only the final computational output	Increases the ability to compute over distributed datasets without revealing original data	Higher computational and communication costs/burdens, and difficult to scale
Homomorphic Encryption	Allows computing over encrypted data to produce results in an encrypted form	Only authorized users can see original and/or computed data	Higher computational cost and time
Zero-Knowledge Proof	Allows one party to prove to another party that a particular statement is true without revealing privacy-sensitive information	Increases ability to validate information without disclosing sensitive information	Cost and scalability
Trusted Execution Environment	Creates a secure, isolated execution environment parallel to the main operating system to process sensitive data	Allows faster secure analytics on data compared to encryption-based techniques	Introduces other ways sensitive data can leak
Federated Learning	Allows multiple entities to collaborate in building an ML model on distributed data without sharing original data	Minimizes data sharing while training a combined model	Various data reconstruction or inference attacks are still possible; require consistency across datasets held by multiple entities

Table 1: Overview of Key Technical Approaches Essential for PPDSA

Platform Solutions

Google launched Analytics Hub in May 2021 as a platform for combining datasets and sharing data, dashboards, and machine learning models. Google also launched Datashare, developed more for financial services and based on Analytics Hub. Databricks announced Delta Sharing on the same day as Google's release. Delta Sharing is an open-source protocol for secure data sharing across organizations. In June 2021, Snowflake opened broad access to its data marketplace, coupled with secure data sharing capabilities.

DataVaults

DataVaults delivers a novel framework and architecture that leverages personal data, coming from diverse sources (sensors, IoT, wearables, data APIs, historical data, social network data, activity trackers, health records, demographic profiles, etc.) to help individuals construct their unified personal data hub, collect at a single point all of their personal data in a secure and trusted manner, and retain ownership and control on what to share and with whom, receiving also compensation for the artefacts they place at the disposal of other third parties.

In turn, third party organisations (companies, public sector, NGOs, etc.) arrive at a position where they can request and get access to tons of personal data, which can complement the ones they already manage and that can be used for generating more efficient, effective and value added services, engaging with individuals into an entirely new way for data sharing, which generates trust and an increased feeling of collaboration, as the data owners (e.g. individuals) become the centre of attention and the most important partner and collaborator of those third parties.

At the core of DataVaults lies the DataVaults personal data value chain which could be seen as a multi-sided and multi-tier ecosystem governed and regulated by smart contracts to safeguard personal data ownership, privacy and usage and attribute value to all entities that generate value within this chain and especially data owners. DataVaults delivers a framework and platform that sets, sustains and mobilises this ever-growing ecosystem for personal data sharing and for enhanced collaboration between those who own data (data owners) and those who seek data (data seekers).

The data value chain that is tackled by DataVaults can be seen as a structure of a trusted cycle, which includes:

- Primary Personal Data Providers (Individuals): This tier includes all the individuals which are generating and collecting their personal data from various services, devices and applications. It is these data which is considered "personal" and constitutes the core data of that is of interest of the DataVaults project.
- Data Seekers (Economic Operators): 1st-tier economic operators, that look for enjoying business intelligence based on Primary Personal Data. In this tier, data seekers (organisations of any type) are able to work on the data of the first tier (primary data) and combine them with other types of data they have to create new datasets or relevant derivatives (insights, reports, etc). 2nd-tier economic operators that provide data and services based on analytics or data that is shared and generated by the 1st-tier economic operators.

DataVaults provides the following services to enable users make the most out of their personal data while having maximum control over them:

- Holistic Personal Data Management: Personal data management services, including collection, mining processing, normalisation, formatting and availability at individuals' personal devices level as well as on secure data vaults on the cloud
- Smart Data Interlinking: Personal data are linked to open, linked as well as proprietary data following Linked Data principles and openly (re-)publishing non-sensitive and business critical information to the LOD community
- Novel Data Security: Cryptography, data anonymisation and privacy preservation, remote attestation and trusted data exchange through the utilisation of TPM technologies between the Personal DataVaults and the DataVaults cloud-based engine
- Privacy Risk Assessment: Methods that offer a "situational awareness" picture to individuals with easy-to-understand privacy metrics, revealing the true risk exposure factor of individuals based on the shared data
- Privacy Preserving and Data Secure Retention Mechanisms: Accommodate the generation of anonymised "digital twins" of individuals, as well as specimen clusters ("persona groups") to empower group analytics that contain valuable insights without violating privacy principles
- Twin-fold Data Brokerage Engine: A data brokerage engine to cater for IPR and data license safeguarding, documenting transactions in a privacy preserving, yet indisputable and unforgeable manner, facilitating compensations schemes with third parties (that support the shift to future monetisation streams) through the instantiation of multi-layer real-time micro-contracts specifically tailored to the needs of data sharing, redistribution and utilisation, constructing a bridge between personal data and industrial data platforms
- Edge and Centralised Analytics: Smart balancing of analytics methods to accommodate Edge Analytics as well as centralised operations depending on the degree of data volume, velocity and variety, always in conjunction with the security and privacy modalities allowed by the individual for each kind of analysis.
- Provision of intuitive analytics, reports, smart dashboards and visualisations tailored to the needs of each stakeholder of the domain, including the individual, as well as generic ones for wider use by any interested organisation and by the public

Concluding Observations

Data has become a critical asset to the success of any organisation in today's digital landscape. But having it is one thing. Extracting relevant insights is something else entirely. To this end, data-sharing has emerged as a key enabler to unlock business value. Data sharing is the process of making the same data resources available to multiple applications, users, or organizations. It includes technologies, practices, legal frameworks, and cultural elements that facilitate secure data access for multiple entities without compromising data integrity. Data sharing improves efficiency within an organization and fosters collaboration with vendors and partners. Awareness of the risks and opportunities of shared data is integral to the process.

The ASEAN Framework for Digital Data Governance provides a basis for facilitating the creation of a vibrant data ecosystem fostering collaboration, interoperability and accelerating data innovation at scale. The author views the Digitalisation and Emerging Technologies pillar within the framework as vital to success at any substantive scale. There are now significant regulatory frameworks around the world that provide a blueprint and/or best practice on how to ensure the correct safeguards are in place. The EU approach provides a starting point of how to maintain interoperability and facilitate

cross border data sharing. Core technologies are now tried and tested and, in many cases, widely deployed to support data integrity, data transparency, data security and privacy. The emergence of platforms eases and reduces the barriers to implementing data sharing at scale as the platforms provide the required safeguards, allow data owners to keep control of their data (sovereignty) and now harness the capability to support data monetisation through the ability to transact data.

About the Author

Dr Ilesh Dattani

Ilesh Dattani is the CTO and Founder of Assentian based in the UK, USA, Ireland and Brazil. Prior to this he was the co-founder and CTO of a UK/US Fintech for 9.5 yrs. prior to its acquisition. He is a Certified Information Security Auditor and a Chartered Engineer. Ilesh has an undergraduate degree in Mathematics from University College London and a Masters in Financial Mathematics and PhD in Machine Learning from Yale in the United States. He has 15 yrs. experience of delivering innovative disruptive solutions in the Financial Services Industry and prior to that he spent 8 yrs. in the Civil Aviation sector in Europe and North America. He is an advisor to and/or board member of Cyber Security start-ups in the UK, USA, Singapore and Australia and mentors' start-ups in accelerators in the same countries (IoT Tribe UK/Singapore, MACH-37-USA). He is a member of the ISO/IEC Joint Technical Committees working on Standards in Information Security and Artificial Intelligence. He is the co-founder of a private family fund based in the USA investing in very early stage, high-risk high-impact start-ups.and he is an investment advisor to Prota Ventures - a US DeepTech Fund. He is an advisor to Enterprise Ireland (an agency of the Irish Government) on the strategic development of their Disruptive Technologies Fund. Finally, he currently supports spin-outs as an investor and CTO focussed on commercialisation of innovative disruptive technologies in the Insurance, Cyber Security, Supply Chain Management and Privacy-Preserving Cryptography areas in the United States, UK, Singapore, Australia and Brazil. He is fluent in English and French and has working proficiency in German and Portuguese.

ASSENTIAN LIMITED

A global Cyber Security and Blockchain Innovation Hub working with clients and partners to address current business challenges by exploring the potential of emerging technologies.

Over 25 Years of Experience in working with emerging technologies like Al, Machine Learning, Internet of Things, Blockchain and others

Support with regulatory compliance around Cyber Security and Data Protection including ISO 27001, PCI DSS, GDPR, LGPD and others

Certified Information Security Auditors, ISO 27001 and CREST Accredited Approved UK Crown Commercial Services Suppliers

Over 15 years' experience in implementing standards and regulatory requirements around information security and data privacy including GDPR.

A clear understanding of Information Security and Data Privacy provides a strong underpinning when looking at potential use cases and defining some of the compliance and ethical considerations that might arise