WHITE PAPER

The Five Data Localization Strategies for Building Data Architectures

How to operationalize data localization requirements in a multi-compute environment

Sophie Stalla-Bourdillon
Senior Privacy Cousel and Legal Engineer, Immuta

Elena Elkina
Aleada Consulting

Chris Ireland
Immuta Scholar

Caroline Jackson
Immuta Scholar

IMMUTA™

April 2022

Table of Contents

Introduction

In recent years, there has been a gradual increase in the number of countries adopting data localization laws.[1] According to a report by the Information Technology & Innovation Foundation, the number of data localization laws has more than doubled since 2017, rising from 35 countries with 67 controls to 62 countries with 144 controls.[2] These data localization mechanisms can apply to both personal data and non-personal data.

Organizations that utilize data across jurisdictions can no longer ignore data localization requirements. These requirements are present in key markets such as China, the European Union, Australia, India and more. Following the Schrems II case, organizations must now comply with what has been termed a European Union (EU) soft-data localization regime.[3]

The policy reasons behind data localization vary across jurisdictions, with some being more legitimate than others. These measures can be intended: (1) to force compliance with data protection laws and improve individual "control" over data, (2) to facilitate data re-use by local businesses and organizations, (3) to facilitate law enforcement's access to data, (4) to make surveillance of citizens easier, and (5) even to offset the impact of any potential sanctions by other countries that might restrict access to data or online services.[4] Therefore, the three main triggers of data localization requirements are privacy and data protection, data sovereignty, and public security concerns.[5]

Data localization is inherently fickle. If implemented too strictly, it becomes self-defeating, undermining system operational robustness by limiting availability and hampering disaster recovery efforts, as well as hindering data sharing and thereby making data science and Artificial Intelligence (AI) research and innovation more challenging.

This white paper proposes a practical and balanced approach to data localization requirements. To this end, it distinguishes between data localization, data anonymization and data segmentation, and combines these techniques with a desire to protect both data and metadata.

While this white paper suggests that data localization requirements should be taken seriously, it does not imply that all types of these requirements are equally legitimate and unchallengeable. In fact, it is essential to identify the regulatory objective behind each data localization requirement prior to elaborating a data strategy.

The white paper is organized as follows: Section one offers an overview of the data localization spectrum and unpacks data localization requirements stemming from India, the EU, and China. Section two describes a typical analytics environment data stack, composed of five key data elements, and highlights data localization implications for each of them. Section three illustrates how data localization requirements can be met in a multi-compute environment through five data localization strategies.

1 Taylor RD, "'Data Localization': The Internet in the Balance" (2020) 44 Telecommunications Policy 8.

2 Cory N, Dascoli L, "How Barriers to Cross-Border Data Flows Are Spreading Globally, What They Cost, and How to Address Them." <https://itif.org/publications/2021/07/19/how-barriers-cross-border-data-flows-are-spreading-globally-what-they-cost> accessed March 28, 2022.

3 Chander A, "Is Data Localization a Solution for Schrems II?" (2020) 23 Journal of International Economic Law 772: It is soft in the sense that data transfers are still permitted but certain controls are required to ensure data subject rights remain unaffected. Furthermore, at the EU legislative level, the proposed Data Governance Act, if enacted, may further restrict the transfer of non-personal data.

4 Cory (n 2).

5 For example, compare the Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC [2016] (General Data Protection Regulation, "GDPR"), the Privacy Act 1988 (Australia), the Cybersecurity and Personal Information Protection Laws (China), Federal Laws 242-GZ and 152-FZ (Russia).

❶ The Data Localization Spectrum

Defined broadly, ***"data localization refers to a mandatory legal or administrative requirement directly or indirectly stipulating that data be stored or processed, exclusively or non-exclusively, within a specified jurisdiction".***[6]

It is very rare under data localization requirements that input data, output data and metadata should all reside within only one jurisdiction. Most jurisdictions allow for an exception-based or conditional transfer system. In these cases, cross-jurisdiction data operations can be permitted for specific purposes and/or if a set of conditions is met.

Data localization laws vary in scope from having no restrictions to imposing overarching restrictions on the transfer of data out of jurisdiction.

- **No restrictions** – The California Consumer Privacy Act 2018 does not restrict the cross-border transfer of data.[7]
- **Mirroring** – Mirroring requires keeping a copy of data within a certain jurisdiction. Under a mirroring requirement, an organization may transfer data outside of the jurisdiction as long as the organization maintains a copy of the data in the original jurisdiction. An example of mirroring arises in the context of the India Company Act.[8] The Companies Act, 2013 requires covered organizations to store financial information at the registered company's office.
- **Soft data localization (or international transfer restrictions)** – The EU General Data Protection Regulation (GDPR) only permits the cross-border transfer of data if appropriate safeguards are in place or the third country has been granted an adequacy decision by the European Commission;[9] The Chinese Personal Information Protection Law restricts the cross-border transfer of personal data unless a set of conditions are met.[10]
- **Hybrid data localization (or mirroring and international transfer restrictions)** – The proposed Indian Data Protection Bill imposes a mirroring requirement as well as transfer conditions for sensitive personal data. It requires that copies of sensitive personal data are always stored within India, even if the data satisfies the legal conditions to be processed in another country. This is more stringent than pure mirroring, which allows relatively unrestricted transfer as long as a copy of the data is stored within the jurisdiction. This is also more stringent than soft data localization.

Let's dive into three jurisdictions to better understand how these data localization requirements are formulated.

6 Svantesson D, Data localisation trends and challenges: Considerations for the review of the Privacy Guidelines (OECD Digital Economy Papers No. 301, OECD Publishing 2020) 8.

7 Determann L and Gupta C, "India's Personal Data Protection Act, 2018: Comparison with the General Data Protection Regulation and the California Consumer Privacy Act of 2018" (2019) 37 Berkley Journal of International Law 511INT'l L. 481 (2019), at p. 511.

8 Sections 88, 92, 94, (Indian) Companies Act 2013.

9 Article 45, GDPR.

10 Article 40, Personal Information Protection Law.

India

Applicable Law

India currently takes a sectoral approach to cross-border transfers, focusing on company financial information, payment systems, and insurance. The Companies Act, 2013 requires covered organizations to store financial information at the registered office of the company.[11] Meanwhile, the Reserve Bank of India has issued a directive requiring any payment systems providers to store payment data in India, with a mirroring requirement for cross-border transactions.[12] Insurers are also required to store insurance data within India under the Insurance Regulatory and Development Authority (IRDAI) (Maintenance of Insurance Records) Regulations, 2015.[13]

A limited category of sensitive personal data is also subject to some general transfer requirements. Under the Information Technology (Reasonable Security Practices and Procedures and Sensitive Personal Data or Information) Rules, 2011 ("the 2011 Rules"), "sensitive" personal data can only be transferred, either within India or abroad, if the data subject has consented or if necessary for a contract between company and data subject. The receiving entity must adhere to the same standard of data protection as the transferring company.[14] Sensitive data is defined in the 2011 Rules as information relating to passwords, financial information, health data including medical history, biometric data, and sexual orientation.[15]

However, India has been considering a more comprehensive approach to data protection. The proposed Data Protection Bill, 2021 ("the DPB"), if brought into force, would restrict the transfer of certain categories of personal data subject to stringent conditions. The bill is expected to be passed by the Indian Parliament in 2022.

The DPB introduces broad data localization requirements for sensitive and critical personal data. Sensitive personal data is a considerably broader category than under the 2011 Rules, and is defined as data revealing, relating to, or constituting an individual's health, finances, official identifiers like a passport or driver's license, sex life, sexual orientation, biometrics, genetics, transgender or intersex status, caste or tribe, and religious and political beliefs.[16] Critical personal data refers to data designated as such by the Central Government.[17]

Under the proposed Bill, sensitive personal data can only be transferred out of India for processing under certain conditions, and in any event should continue to be stored in India.[18] There are three situations in which transfer can occur:

1. In pursuance of a contract or inter-group scheme approved by the Data Protection Authority (DPA) in consultation with the Central Government
2. When the transfer is to a country, entity, or class of entities whom the Central Government has approved
3. When the DPA, after consulting with the Government, agrees that a transfer is necessary for a given purpose.

In addition, the data subject's consent is always required.[19]

11 Sections 88, 92, 94, (Indian) Companies Act 2013; Companies Accounts Rules 2014.

12 RBI Directive 2017-18/153, para 2(i); Payments and Settlements Act 2007.

13 IRDAI (Maintenance of Insurance Records) Regulation 2015, para 3(9).

14 Section 7, Information Technology (Reasonable Security Practices and Procedures and Sensitive Personal Data or Information) Rules 2011.

15 Section 3, Information Technology (Reasonable Security Practices and Procedures and Sensitive Personal Data or Information) Rules 2011.

16 The Central Government is also entitled to set new categories of sensitive data; Section 3(41), Data Protection Bill 2021.

17 As of writing, the Central Government has not designated any types of data as critical. Section 3(41), Data Protection Bill 2021.

18 Section 33(1), Data Protection Bill 2021.

19 Section 34(1), Data Protection Bill 2021.

Critical personal data can only be processed in India and may only be transferred outside of India to (1) countries or entities whom the Government has approved, or (2) to healthcare professionals in the case of medical emergencies.[20]

Another draft Bill under consideration is the Digital Information Security in Healthcare Act 2018 (DISHA), which if passed, would give data subjects ("owners") the right to require explicit permission each time their data is transferred in an identifiable form.[21]

Takeaway

Transfer of data outside of India can currently be considered largely unrestricted. If an organization falls within the scope of current data localization requirements, it may still process data abroad as long as an up-to-date version of the data is also maintained in India.

If the DPB is enacted, cross-border transfers of sensitive and critical personal data will become more onerous. Under the DPB, both business-to-business and consumer-to-consumer data will likely be treated in the same way.[22] Therefore, organizations need to be aware that any data deemed to be sensitive or critical personal data will fall under the scope of the DPB, even if it has not been collected in a business-to-consumer relationship. Non-sensitive personal data will continue to be transferred without restriction.

It also appears that localization requirements will apply to sensitive and critical data even if it is anonymized. In addition to the DPB, India is considering a Non-Personal Data Governance Framework ("the NPDGF"), which would include anonymized data. The latest report on the NPDGF proposes that anonymized data should inherit the sensitivity of the underlying personal data and be subject to the same storage requirements.[23] Organizations should be aware that if the NPDGF is passed, anonymized sensitive or critical personal data will likely be subject to the same localization requirements as if it were non-anonymized. De-identification, including pseudonymization, will therefore not prevent localization requirements from applying to sensitive and critical data.[24]

In any case, setting policies upon data that have an impact on accessibility (who can access the data) and visibility (what the data looks like to data users), such as such de-identification policies, can always happen outside of India.

20 Sections 33(2); 34(2), Data Protection Bill 2021.

21 Section 28(8), DISHA 2018.

22 There are a number of exceptions listed in Chapter 8 of the Data Protection Bill 2021, but the exemption list does not include business-to-business relationships.

23 Draft Report by the Committee of Experts on Non-Personal Data Governance Framework (2020), para 8.15 <https://static.mygov.in/rest/s3fs-public/mygov_160975438978977151.pdf> accessed March 28, 2022.

24 De-identification is considered to be a form of security safeguard that does not alter the status of the underlying data; Section 24(1), Data Protection Bill 2021.

European Union

Applicable Law

In the EU, there are two layers of data localization requirements to consider: those at the EU level (applicable to all 27 Member States), and those enacted by individual countries.

While the EU's General Data Protection Regulation (GDPR) does not contain explicit data localization requirements in that data is not required to be kept within the EU, the GDPR does introduce data transfer requirements by imposing restrictions upon the international transfer of personal data to third countries. For the purposes of this paper, we refer to the GDPR's data transfer requirements as a type of soft data localization requirement because of their impact; if EU data cannot be transferred outside the EU in certain circumstances, then the data transfer requirement in effect becomes a data localization requirement. Chapter 5 of the GDPR outlines the requirement that a regular cross-border data transfer must generally be either (1) on the basis of an adequacy decision by the European Commission,[25] or (2) subject to appropriate safeguards.[26] These provisions are designed to ensure that the level of protection afforded by the GDPR is not undermined by a transfer to a third country.[27] The European Data Protection Board has recently clarified that transfer restrictions do not apply to cross-border data movements between employees of an organization subject to GDPR.[28] The same is true for cross-border data movements between controllers/processors and "mere" data subjects.[29]

The decision of the Court of Justice of the EU in the Schrems II case means that the adequacy decision grounding the Privacy Shield[30] framework for international transfers to the United States is no longer valid, although a new adequacy decision is expected soon.[31] In the meantime, Standard Contractual Clauses eventually complemented by technical measures and derogations under GDPR Article 49, remain valid alternatives. Business-to-business data like employee or sales-related information is often considered to have a low risk of access requests by intelligence agencies.[32]

At a national level, at least 21 of the 27 EU Member States have introduced data localization measures to information concerning things like taxation, company data, defense, national security and health data.[33] These measures involve further access controls and can be justified by the high sensitivity of the data.

25 Article 45, GDPR.

26 Article 46, GDPR. Article 49, GDPR provides some specific derogations from Article 45 and 46 that can sometimes be relied upon.

27 Article 44, GDPR.

28 European Data Protection Board, Guidelines 05/2021 on the Interplay between the application of Article 3 and the provisions on international transfers as per Chapter V of the GDPR (2021), at p. 6-7.

29 Ibid. at p. 5-6.

30 Commission Implementing Decision (EU) 2016/1250 of 12 July 2016 pursuant to Directive 95/46/EC of the European Parliament and of the Council on the adequacy of the protection provided by the EU-U.S. Privacy Shield [2016]; Case C-311/18 Data Protection Commissioner v Facebook Ireland Limited and Maximillian Schrems [2020] EU:C:2020:559.

31 See European Commission and United States Joint Statement on Trans-Atlantic Data Privacy Framework, March 25, 2022 <https://ec.europa.eu/commission/presscorner/detail/en/ip_22_2087> accessed March 28, 2022.

32 Dutch Supervisory Authority, Data Protection Impact Assessment on Microsoft Teams, OneDrive, Sharepoint and Azure AD, Version 1.1, February 16, 2022, <https://www.rijksoverheid.nl/binaries/rijksoverheid/documenten/publicaties/2022/02/21/public-dpia-teams-onedrive-sharepoint-and-azure-ad/Public+DPIA+Teams+OneDrive+SharePoint+and+Azure+AD+16+Feb+2022.pdf > accessed February 25, 2022; See also US Department of Commerce, Information on U.S. Privacy Safeguards Relevant to SCCs and Other EU Legal Bases for EU-U.S. Data Transfers after Schrems II (2020). <https://www.commerce.gov/sites/default/files/2020-09/SCCsWhitePaperFORMATTEDFINAL508COMPLIANT.PDF> accessed February 26, 2022.

33 Ursic H, Nurullaev R, Cuevas MO and Szulewski P, "Data Localisation Measures and Their Impacts on Data Science" Handbook on Data Science and Law (Edward Elgar 2018); European Commission, Commission Staff Working Document on the free flow of data and emerging issues of the European data economy accompanying the document Communication Building a European data economy (European Commission, 2017) 9.

Furthermore, the European Commission's Proposal for a Regulation on European data governance (Data Governance Act)[34] would introduce a regime of shared European "data spaces" covering sectors where data mining could unlock significant insight, such as healthcare and energy. The Commission's proposal would also introduce restrictions upon the transfer of non-personal data related to these sectors to third countries.[35] The rationale behind the extension of transfer restrictions to non-personal data is to enable the EU to realize the value from its citizens' data.

Takeaway

The European Union has introduced deliberate measures to make the cross-border transfer of personal data more burdensome. Therefore, organizations that collect and process the personal data of individuals accessing their goods and services from the European Union should adopt a cautious approach to cross-border transfers. There are no exceptions for the collection and processing of personal data in the context of business-to-business relationships.

Pseudonymization, if implemented in line with the European Data Protection Board's recommendations,[36] should make the establishment of a legal ground for transfer much easier. Arguably, if the data is anonymized, no restriction applies to its transfer.[37]

In any case, just like under the Indian framework, setting policies upon data with an impact upon accessibility and visibility, such as such de-identification policies, can in principle happen outside of the EU as long as the latter does not imply any restricted data movement.

34 Proposal for a regulation of the European Parliament and of the Council on European data governance (Data Governance Act) [2020]. See also Proposal for a regulation of the European Parliament and of the Council on harmonised rules on fair access to and use of data (Data Act) [2022].

35 Retical 19 and Article 5(11), Data Governance Act; recently accepted by Ambassadors to the EU <https://www.euractiv.com/section/data-protection/news/eu-counties-green-light-new-data-governance-framework/>accessed February 26, 2022.

36 European Data Protection Board, Recommendations 01/2020 on measures that supplement transfer tools to ensure compliance with the EU level of protection of personal data, 2020. For a discussion about the upshot of these recommendations see Stalla-Bourdillon S, Rossi A, "The Technical Fix for International Data Transfers: A Word of Caution" (Immuta 2021) <https://www.immuta.com/downloads/the-technical-fix-for-international-data-transfers-a-word-of-caution/> accessed February 22, 2022.

37 Recital 26 & Article 44 GDPR; "International Transfers after the UK Exit from the EU Implementation Period" (ICO). <https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/international-transfers-after-uk-exit/> accessed March 28, 2022.

China

Applicable Law

The legal framework for data localization in China stems from three laws: (1) the 2017 Cybersecurity Law,[38] (2) the 2021 Data Security Law,[39] and (3) the 2021 Personal Information Protection Law.[40]

The most recent Personal Information Protection Law reflects the Chinese Government's increased focus on the commercialization of citizens' personal data and the subsequent power available to companies that hoard data.[41]

There are four key definitions that need to be considered when interpreting China's data localization regime: personal information,[42] personal information handler,[43] important data[44] and critical infrastructure operators (CIIOs).[45] Unlike other regimes, there is no distinction between low and high risk categories of data.

The Personal Information Protection law has extraterritorial effect and applies to foreign companies that: (1) market goods and services in China or (2) analyze the data of individuals residing within China.[46] It does not apply to personal data that has undergone an anonymization process.[47] The Personal Information Law introduces the concept of de-identification, which is similar to the EU concept of pseudonymization.[48] It is important to note that de-identification does not reduce the compliance requirements for personal data. The Chinese law classifies it as a technical security measure to prevent unauthorized access to data and describes the process as personal data "undergoing handling to ensure it is impossible to identify specific natural persons without the support of additional information."[49]

For any transfer, a self-assessment process must be undertaken, where organizations are required to consider: (1) the legality and purpose of the transfer, (2) the quantity, scope, categories and degree of sensitivity of the data being transferred, (3) any risks to national and rights of individuals and organizations, (4) the possibility of data leaks and distortion during transfer, (5) the responsibilities and security capabilities of the foreign receiving party, (6) the possibility of data leaks and distortion after transfer, and (6) the responsibilities and duties outlined in the transfer agreement.[50]

38 Cybersecurity Law (China); Translation: available at <https://www.newamerica.org/cybersecurity-initiative/digichina/blog/translation-cybersecurity-law-peoples-republic-china/> accessed March 28, 2022.

39 Data Security Law (China); Translation: available at <https://digichina.stanford.edu/work/translation-data-security-law-of-the-peoples-republic-of-china/> accessed March 28, 2022.

40 Personal Information Protection Law (China); Translation: available at <https://digichina.stanford.edu/work/translation-personal-information-protection-law-of-the-peoples-republic-of-china-effective-nov-1-2021> accessed March 28, 2022.
For a more in-depth overview of Chinese law, see Dorwart D, "Demystifiying Data Localization in China: A Practical Guide" (February 2022) <https://fpf.org/wp-content/uploads/2022/02/Demystifying-Data-Localization-Report.pdf> accessed March 28, 2022.

41 AP NEWS, "New Chinese Law Tightens Control over Company Data on Users" (August 20, 2021) <https://apnews.com/article/technology-business-china-data-privacy-1d3fcbac4549c6968c07897900c96cc3> accessed March 28, 2022.

42 Defined as any information that can identify or purports to identify a natural person; Article 76(5), Cybersecurity Law.

43 Defined as an organization that decides what data is processed, and how it is processed; Article 73(1), Personal Information Protection Law.

44 Referenced in both the Cybersecurity Law and Data Security Law. There is no clear definition; the Data Security Law requires that, under Article 21, that central and regional government departments catalog what falls under the classification.

45 Bodies that operate "critical information infrastructure which—if destroyed, suffering a loss of function, or experiencing leakage of data—might seriously endanger national security, national welfare, the people's livelihood, or the public interest." Article 31, Cybersecurity Law.

46 Article 3, Personal Information Protection Law.

47 Articles 4 and 73, Personal Information Protection Law.

48 Article 73, Personal Information Protection Law.

49 Articles 51 and 73, Personal Information Protection Law.

50 Article 5, Outbound Data Transfer Security Assessment Measures (China); Translation at <https://digichina.stanford.edu/work/translation-outbound-data-transfer-security-assessment-measures-draft-for-comment-oct-2021/> accessed March 28, 2022.

In addition to self-assessment and, in the case of personal data, obtaining consent and signing contracts with foreign recipients, organizations are also required to either:[51] (1) apply for a security assessment with their local provincial level cybersecurity department,[52] (2) undergo a certification program, (3) use a standard form contract drafted by the Cyberspace Administration, or (4) adhere to other, yet to be determined mechanisms.[53] The Cyberspace Administration is likely to request industry standard technical and organizational controls to ensure that the data can only be accessed for its specified purpose.

A security assessment is required when CIIOs and any personal information handlers (together referred to as "data handlers" by Chinese law) either: (1) handle important data, (2) store the personal data of more than one million people, or (3) wish to transfer the personal data of more than 100,000 people or the sensitive personal information of more than 10,000 people apply for a security assessment before transferring it out of China.[54] Under these circumstances, cross-border flows appear to require explicit permission from the regulator, an example of hard data localization.[55] Generally, small operators will be able to use contractual clauses or certification as a basis for transfer. The certification process is yet to be outlined by the Chinese government.[56] Large operators that meet the requirements outlined above will be required to apply for an assessment. The assessment is carried out by the Chinese Cyberspace Authority and, following the first assessment, there is an annual reporting obligation.[57]

Regulations suggest that important data is that which, if tampered with, sabotaged, leaked, illegally acquired, or illegally used, may cause harm to national security or the public interest.[58] The mechanism for the transfer of important data requires that organizations be subject to detailed monitoring and auditing by state authorities.[59] Organizations are required, as part of the security assessment process, to demonstrate that there is a "business requirement [that] it is truly necessary to provide it outside [of Mainland China]."[60] Further detail on the assessment process is expected in 2022.

In practice, this means that most organizations will be required to both process and store data within China unless they have obtained consent from a data subject or the regulator.[61] This restriction applies regardless of the nature of the personal data, and could include data collected, relating to identifiable natural persons, in business-to-business relationships.

51 Article 40, Personal Information Protection Law.

52 Article 4, Outbound Data Transfer Security Assessment Measures.

53 Article 38, Personal Information Protection Law.

54 Article 4, Outbound Data Transfer Security Assessment Measures.

55 Article 11, Outbound Data Transfer Security Assessment Measures.

56 It is expected that further information on the certification process, including the bodies that can perform the process will be published in 2022.

57 Article 40, Regulations on Network Data Security Management.

58 Article 73(3), Regulations on Network Data Security Management; Article 9, Regulations on Critical Information Infrastructure Security Protections (China); Translation available <https://www.chinalawtranslate.com/en/Regulations-on-Critical-Information-Infrastructure-Security-Protections/> accessed March 28, 2022.

59 Articles 38 and 39, Cybersecurity Law.

60 Article 37, Cybersecurity Law. It is not clear whether the threshold for a requirement to be "truly necessary" is the same as the wording "need to provide" in Article 40 of the Personal Information Protection Law.

61 Article 39, Personal Information Protection Law.

Takeaway

China has also introduced deliberate measures to make the cross-border transfer of data more burdensome, but goes further than the EU in that it also restricts the transfer of important data. Once again, organizations that collect and process the personal data of individuals accessing their goods and services from China should thus adopt a cautious approach to cross-border transfers. There does not seem to be any exceptions for the collection and processing of personal data in the context of business-to-business relationships.

Arguably, if the data is anonymized (rather than de-identified), no restrictions apply to its transfer. Furthermore, if a data processor is located outside of China and needs to transfer data for the purposes of providing goods or services that a Chinese citizen has solicited, no certification or self-assessment is necessary.[62]

With this said, nothing seems to preclude setting policies like de-identification upon regulated data with an impact upon accessibility and visibility from outside China, as long as the latter does not imply any restricted data movement.

62 Article 35, Regulations on Network Data Security Management.

❷ The Data Stack

This section describes five key data elements that comprise the foundational data stack of any analytics environment, which should be considered when implementing data localization requirements. The term metadata, or information about other data, is used to distinguish elements that describe analytics data in the data stack. This can include the context, structure, semantic meaning, usage, sources, and storage of analytics data. Metadata therefore includes data dictionaries, user data, and audit logs.

It is possible to represent each data element in relation to their content (whether they comprise personal data or not) and their risk level (whether they create low or medium/high risks for the individual to whom the data pertain).

	PERSONAL DATA	NON-PERSONAL DATA
Low Risks	▪ User Data ▪ Audit Logs	▪ Data Dictionary (when membership inference is not a concern)
Medium/ High Risks	▪ Analytics Data ▪ Query Results	▪ Data Dictionary (when membership inference is a concern)

Table 1. The five data elements and their related risk levels

Before analyzing each data element in detail, we should distinguish three types of techniques used to protect data elements and comply with restrictions imposed upon international transfers, as illustrated in Figure 1:

- Data localization means that the data should live within the local jurisdiction in which its subject (or object) sits. It is considered absolute when the data never leaves the jurisdiction in which it lives, even temporarily, and relative when a predetermined set of movements out of the jurisdiction in which the data lives is allowed. As explained in section one, absolute data localization is rarely the applicable standard.[63]

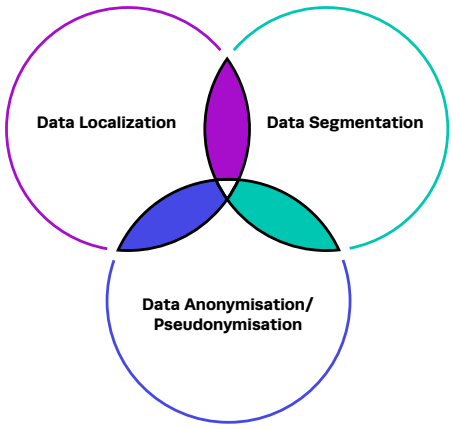

Figure 1: Data protection techniques

63 See for example the Information Commissioner's Office in the UK, who distinguishes between restricted transfer and unrestricted transit: "Transfer does not mean the same as transit. If personal data is just electronically routed through a non-UK country but the transfer is actually from one UK organisation to another, then it is not a restricted transfer." Information Commissioner's Office, "Guide to the General Data Protection Regulation (GDPR)." <https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/international-transfers-after-uk-exit/#whatarethe> accessed March 28, 2022.

- Data pseudonymization/anonymization means that the data should undergo a de-identification process in order to reduce re-identification risks. It is only when the re-identification risk is considered reasonably remote that the data will be deemed anonymized. Although anonymization techniques are maturing across the globe and converging, each jurisdiction has its own legal test for determining whether the data is outside the scope of the framework.[64] Data pseudonymization/anonymization can help achieve relative data localization.
- Data segmentation means that access control is applied on the data so that data is rendered inaccessible to certain data user groups or data users with certain attributes, e.g., data users located in certain jurisdictions. Data segmentation can help achieve relative data localization.

Analytics Data

This is data used for analytics purposes. In a data science world, data is analyzed to derive metrics and key performance indicators (KPIs), enhance business outcomes, and drive business processes. In many instances, analytics data can consist of personal data like customer or patient records or website visitor activities. There are direct commercial gains related to the exploitation of this data, so it is understandable that some countries are taking steps to ensure that any economic benefits support the country where the initial data was collected.

CUSTOMER ID	TRANSACTION TIMESTAMP	TRANSACTION VALUE
a2eb9dce-3640-11ec-8d3d-0242ac130003	2021-10-26T09:40:01+00:00	3234.33
a70deede-3640-11ec-8d3d-0242ac130003	2021-10-27T01:40:01+00:00	233.20
aa61ae18-3640-11ec-8d3d-0242ac130003	2021-10-30T09:44:02+00:00	434.55

Table 2. Structured data

Within an analytics environment, data is usually structured in relational tables (see Table 2), semi-structured (i.e. to conform to some standard such as JSON, as shown in Figure 2[65]), or unstructured (i.e. free text or images). Managing all these types of data to derive insights is usually a key requirement for data teams.

64 To know more about pseudonymization, de-identification and anonymization see Stalla-Bourdillon S, Jonnalagadda S, "A Data Protection Grammar – Navigating Regulatory Landscape for Compliance in Multi-Compute Environment" (Immuta 2021). <https://www.immuta.com/downloads/a-data-protection-grammar/> accessed March 28, 2022.

65 It is possible to query semi-structured data directly through SQL, load the data into a relational table, and derive a schema, just like for a traditional relational table.

```
SELECT raw:store.transaction FROM store_data
+-----------------------------------------------------------+
| transaction                          |
+-----------------------------------------------------------+
| {                                |
|  "customer_id":"a2eb9dce-3640-11ec-8d3d-0242ac130003", |
|  "timestamp":"2021-10-26T09:40:01+00:00",
|  "value":3234.33,                    |
| }                                |
+-----------------------------------------------------------+
```

Figure 2. Semi-structured data

Data localization restrictions on analytics data introduce significant challenges for the field of data science.[66] For instance, effective machine learning models require high volumes of quality data. The more quality data available, the easier it is for a model to offer useful insight into a particular problem. Restricted global transfers of data could mean more input data quality and representativeness challenges for data scientists building artificial intelligence systems.

Generalization requires that a model is useful at rendering meaningful predictions using sample data. This is a central assumption underlying the development of any predictive model. Restricted global transfers can introduce bias in a model, particularly if outcomes are not universal. For example, if fraud patterns are different in China than in Denmark, a model developed on data from Denmark will not provide meaningful results when applied to events in China.

When identifying data items are included within analytics data, the latter falls within the remit of data localization requirements under most regimes. **Data localization techniques will thus have to be implemented to guarantee that the analytics data lives within the local jurisdiction in which its subject (or object) sits, unless an exception applies.**

Data Dictionaries

A data dictionary is an organized set of attribute (sometimes referred to as columns) names, descriptors, types, validation rules and, if applicable, relationships with other attributes. Attribute names and descriptors are usually collected in tables. Mentioned in Section 2.1, data dictionaries are relevant for both structured and semi-structured data. Data dictionaries can be complemented by data source statistics to monitor changes to data sources, like through schema monitoring.

In a table with health record information (see Table 3), one of the attributes listed might be an individual's social security number. The data dictionary would include its attribute name (SSN), descriptor ('this is an individual's government issued social security number'), type (text), validation rules (a regular expression), and its relationship with other attributes (n/a).

66 Ursic (n 33).

NAME	DESCRIPTOR	TYPE	VALIDATION RULES	RELATIONSHIPS
PROVIDER_ID	A unique identifier for the provider	Text	Not NULL and Unique	n/a
SSN	This is an individual's government issued social security numbers	Text	RegEx('^(?!666|000|9\\d{2}\\d{3}-(?!00)\\d{2}-(?!0{4})\\d{4}$')	n/a
MD	This is an individual's medical doctor ID	Number		Many to one relationship with attribute 'ID' in table medical_doctors

Table 3. A data dictionary

Data dictionaries are defined when a data set is created either with the database itself, or in the application code which then modifies the database through a Extract, Transform, Load, and Govern (ETLG) process. This means that the data dictionary definition will co-exist in two different locations, one acting as the master (the application code) and the other which is a replica of the dictionary (the database).

Personal information is most likely to occur in a data dictionary in the attribute descriptors or names of individuals. From a data redundancy standpoint, this would be bad practice and is therefore rarely seen, as user information and attributes would then be repeated in the database. Data redundancy can lead to databases taking up more secondary storage space than necessary, as well as introduce the risk that the same information might differ within the same database.

It is thus unlikely that data dictionaries will fall within the remit of data localization requirements. Data dictionaries do not relate to a singled-out individual. Mirroring data dictionaries, making them live in another jurisdiction, should therefore not be an issue. This does not mean that access to data dictionaries should never be considered sensitive, especially when the concern is whether or not an individual is a member of a table (membership inference). When this is the case, making the data dictionary inaccessible by default to a range of data users remains a valid option. In other words, **data segmentation makes it possible to protect data dictionaries if needed.**

Query Results

Query results are the data that is returned from a database following a Create, Read, Update, or Delete (CRUD) operation. This can take the form of partial or complete data in a single or series of records or a confirmation that a record has been created/deleted. The latter would be an indicator of the number of rows updated, and therefore not be considered personal data. For examples of query results see Figure 3 and Table 4.


```
[{
  "id": 7,
  "timestamp": "2021-10-26T09:40:01+00:00",
  "value": 3234.33
 },
 {
  "id": 8,
  "timestamp": "2021-10-27T01:40:01+00:00",
  "value": 233.20
 },
 {
  "id": 9,
  "timestamp": "2021-10-30T09:44:02+00:00",
  "value": 434.55
 }
]
```


Figure 3. Returned results in JSON format

ID	CUSTOMER ID	TRANSACTION TIMESTAMP	TRANSACTION VALUE
7	a2eb9dce-3640-11ec-8d3d-0242ac130003	2021-10-26T09:40:01+00:00	3234.33
8	a70deede-3640-11ec-8d3d-0242ac130003	2021-10-27T01:40:01+00:00	233.20
9	aa61ae18-3640-11ec-8d3d-0242ac130003	2021-10-30T09:44:02+00:00	434.55

Table 4. Returned results presented in a table (with the customer ID not selected)

Depending on the data set that has been queried, and the attributes that have been selected, it is certainly possible for a query result to contain personal data. This is still possible if it is only stored temporarily in system memory for the purposes of a processing operation.

For example, if a query returns a record containing a complete SSN attribute, it will be stored in system memory whilst being processed by an application. If the application and database are in separate jurisdictions, then a cross-border transfer of data has taken place, barring any exceptions or limitations.

One method used to ensure that individuals are not identified during aggregate processing is differential privacy, which can restrict acceptable queries to those that are not specific enough to identify characteristics about a single individual. Ultimately, if query results are considered to be anonymized, most restrictions upon transfer under privacy and data protection frameworks can be lifted. However, it's likely that fine-grained access control will remain a key context control to boost the anonymization claim.[67]

If there is a need to query the analytics data from a different jurisdiction than the one in which it usually lives, then data localization requirements are almost certainly unavoidable. When querying medium and high-risk analytics data, **coupling data pseudonymization/anonymization with data segmentation should make it possible to achieve a high level of protection.**

2.4 User Data

User data, as illustrated in Figure 4, is information that relates to individuals that query the analytics data and have access to the data dictionaries and analytics data. This typically includes personal data, although the categories of subjects to whom the data relates is certainly different from those represented within the analytics data. User data typically includes user properties such as user identity or name, email addresses, user attributes, and groups.

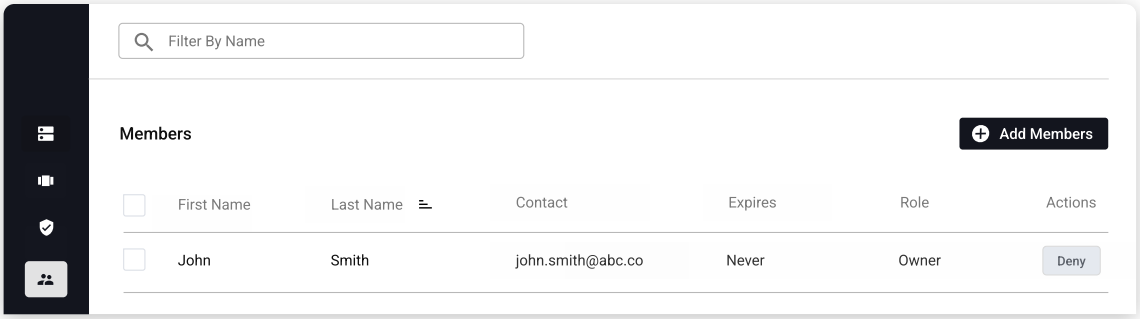

Figure 4. User data

In our healthcare records example, user data would not be included in the relational table. It would usually be captured in a separate database, typically an identity management system or a metadata database.

As long as user data does not involve sensitive attributes, it is usually considered low-risk personal data. This should make it relatively easy to find a justification for transferring the data to another jurisdiction or granting access to data users located in another jurisdiction. It is unlikely that user data constitutes important data under Chinese law or that intelligence services will be after that data, especially if this data is available through other means like organizations' websites or social media platforms. This does not mean, however, that user data should be widely accessible across all jurisdictions. Making user data live within the jurisdiction that the data user and data governance control plane sit for permissioning purposes makes more sense as data movements are reduced to a minimum. In other words, **data segmentation is a valid option for protecting user data.**

67 For an overview of how to combine data and context controls, see Stalla-Bourdillon S & Rossi A, "Aggregation, Synthesisation, and Anonymisation: A Call for a Risk-Based Assessment of Anonymisation Approaches" (Immuta 2020) <https://www.immuta.com/downloads/aggregation-synthesisation-and-anonymisation-a-call-for-a-risk-based-assessment-of-anonymisation-approaches/> accessed March 28, 2022.

User attributes should only fall within the remit of data localization requirements when they relate to a specific individual. Processing user properties at the group level, e.g., for policy authoring, should not be impacted by data localization requirements.

Audit Logs

Audit logs are records that contain information about an event that has occurred in a system, be it a particular operation like a file download, or session activity like a user log in. They are necessary to monitor system usage, observing who is accessing data, for which purpose, at what time, and for how long. The earlier the anomalies are detected, the easier it is to implement corrective measures. Generating and managing audit logs is necessary for many information security certifications.

TIMESTAMP	USER ID	USER EMAIL	IP ADDRESS	RESOURCE	HEADERS
2021-10-26T06:54:42+00:00	53e3be16-362a-11ec-8d3d-0242ac130003	example@domain.com	192.168.16.231	/user/profile/change-details	POST /user/profile/update-details HTTP/1.1 Host: domai n.com Content-Type: application/x-www-form-urlencoded Content-Length: 25 address="123 Acme Street"

Table 5. Audit logs

Audit logs can thus include different types of user properties such as user ID, user email, timestamp, and IP address of the device they used to access the data. If kept in the clear, these user properties should usually be considered personal data under most privacy and data protection frameworks. This said, audit logs tend to partially overlap with user data, and therefore should be considered low-risk data in most cases.

Whilst it is possible to try and reduce compliance burden by not transferring these logs across jurisdictions, this can undermine information security. Without a centralized location for all logs, it is difficult to spot wider trends and draw patterns from usage across the network. A sensible approach would be to keep a central repository which would replicate logs from across the network. Data segmentation thus offers a reasonable approach to the protection of audit logs.

❸ The Five Data Localization Strategies

To fully grasp the potential impact of data localization requirements within analytics environments, it is necessary to distinguish between three layers within the technology stack: the storage layer where datastores sit, the compute layer where the analytics platform sits, and the policy layer where the data governance control plane sits.

A complex data analytics environment will usually involve several datastores hosted by different cloud providers or on premises services. Each datastore will usually have its own data analytics platform. The policy layer is not necessarily specific to each datastore or analytics platform, but can be unified and shared across them.

Within each tech stack, a certain set of data elements will be stored and processed. Three design goals are integral to build secure and resilient architectures for analytics environments:

1. **Data replication:** data elements that are essential to the security and resilience of the environment, such as audit logs and commercially-sensitive data, should not live within only one environment component
2. **Data movement minimization:** movements of data elements between layers should be minimized
3. **Data access observability:** access to data should be systematically monitored and regularly audited with the ability to oversee the entire environment all at once if needed

Taking into account both the data stack explored in our second post and the three layers of the tech stack mentioned above, it is possible to identify three data localization levels and derive five data localization strategies. These localization levels include 'Mirroring,' 'Conditional Access,' and 'Restricted Access.' Mirroring is the least restrictive strategy to data re-use and sharing across jurisdictions, with Restricted Access being the most restrictive. Each data localization strategy is a variation of the high-level architecture diagram represented in Figure 5.

STRONG DATA LOCALIZATION	Analytics Data generated within a jurisdiction & Query Results never leave this jurisdiction.
MEDIUM DATA LOCALIZATION	Analytics Data generated within a jurisdiction & Query Results only leave this jurisdiction when an exception applies.
WEAK DATA LOCALIZATION	A copy of the analytics data generated within jurisdiction X always stays in this jurisdiction.

Table 6. The three levels of data localization.

Each strategy precisely defines where each element of the data stack should live, either in one or more datastores, in one or more analytics platforms, or in the data governance control plane. Notably, the most restrictive strategy, 'Restricted Access', is not necessarily the most secure and resilient, as it scores low on data access observability. It is also the least innovation-friendly, as it blocks data sharing across jurisdictions. On the other hand, all 'Conditional Access' strategies make it possible to control the flow of analytics data to outside jurisdictions and to authorize flows with jurisdictions that offer an equivalent level of protection, as access is granted on an exception-by-exception basis.

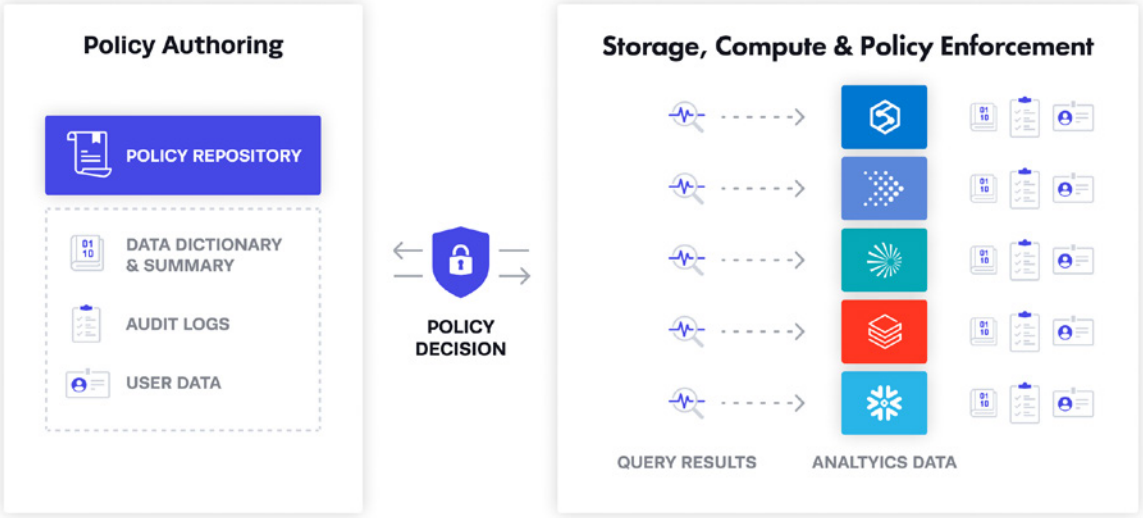

Figure 5. High-level architecture diagram

'Mirroring' maintains a copy of all categories of personal data within the jurisdiction where the data is collected or generated, such as analytics data, user data, and audit logs. Mirroring works well for jurisdictions following the Indian Company Act model.

'Conditional access_low' aims at allowing access to analytics data from outside jurisdictions for specific purposes only. User data and audit logs can be accessed from the jurisdiction in which the data users sit, as well as from the jurisdiction in which the policy layer sits.

'Conditional access_medium' aims at allowing access to analytics data from outside jurisdictions on the condition that both a legitimate purpose is specified and the query results are pseudonymized. User data and audit logs can be accessed from the jurisdiction in which the data users sit as well as from the jurisdiction in which the policy layer sits.

'Conditional access_high' aims at allowing access to analytics data from outside jurisdictions on the condition that both a legitimate purpose is specified and the query results are anonymized. User data and audit logs can be accessed from the jurisdiction in which the data users sit as well as from the jurisdiction in which the policy layer sits.

Conditional access strategies are relevant for both sectorial and horizontal restrictions set upon data transfers, such as restrictions stemming from the EU GDPR or the Chinese Personal Information Protection Law.

Finally, 'Restricted Access' aims at maintaining all data elements within the jurisdiction in which the data is collected or generated. No sharing of data is allowed, although it remains possible to set policies globally. This strategy is sometimes favored by organizations with strict business confidentiality requirements.

Ultimately, this table shows that not all of the data elements of the data stack are necessarily protected in the same way. While data localization frameworks clearly mandate localizing analytics data and query results by default, query results can be protected through pseudonymization/anonymization to avoid localization on a case-by-case basis. In addition, data dictionaries, user data, and audit logs can also be protected and made by default inaccessible through data segmentation to govern access from analytics platforms sitting outside the jurisdiction in which the analytics data usually lives.

When building analytics environments, platform architects and DataOps teams should be sure to carefully select the right data protection technique, data localization, data pseudonymization/anonymization, or data segmentation for each element of the data stack.

	ANALYTICS DATA	QUERY RESULTS	DATA DICTIONARY	USER DATA	AUDIT LOGS
Strategy #1 (Mirroring)	Local datastore(s) only	All data analytics platforms	–Local analytics platform –Data governance control plane –All other analytics platforms	Data governance control plane & local analytics platform	Data governance control plane & local analytics platform
Strategy #2 (Conditional access_low)	Local datastore(s) only	Purpose-based access to query results available to all data analytics platforms	–Local analytics platform –Data governance control plane –All other analytics platforms	Data governance control plane & local analytics platform	Data governance control plane & local analytics platform
Strategy #3 (Conditional access_ medium)	Local datastore(s) only	Purpose-based access to query results & pseudonymized query results available to all data analytics platforms	–Local analytics platform –Data governance control plane –All other analytics platforms	Data governance control plane & local analytics platform	Data governance control plane & local analytics platform
Strategy #4 (Conditional_ high)	Local datastore(s) only	Purpose-based access to query results & anonymized query results available to all data analytics platforms	–Local analytics platform –Data governance control plane –All other analytics platforms	Data governance control plane & local analytics platform	Data governance control plane & local analytics platform
Strategy #5 (Restricted access)	Local datastore(s) only	Local analytics platform only	Local analytics platform only	Local analytics platform only	Local analytics platform only

Table 7. Audit logs

Conclusion

To conclude, adapting data architectures to meet data localization requirements is not an easy task and does imply some tradeoffs. To preserve system operational robustness as well as minimize data movements within a multi-compute environment, it makes sense to abstract key functional layers, i.e., storage, computation, and governance, to create one global repository for the metadata and ensure regulated elements of the data stack never leave local data analytics environments unless a valid ground for transfer has been established and/or appropriate safeguards are in place, in particular to mitigate re-identification risks. The data localization strategy chosen by a data team will ultimately depend upon the extent to which query results should be impacted by privacy or confidentiality preserving techniques.

Learn more about how to implement data access control frameworks built for the data localization laws that impact your organization. Schedule a briefing with Immuta's team of experts.

REQUEST DEMO

IMMUTA™

25 Thomson Place, 4th Floor, Boston, MA 02210 | immuta.com | (800) 655-0982