Background

With more and more organizations delivering digital services through cloud, the information systems are perceived to be centrally deployed and managed. However, emerging data security restrictions and compliance requirements around handling of personally identifiable data in IT systems is affecting the systems design significantly. Many countries have been enacting legislation that requires personal data be kept within geographic boundaries. These regulations have resulted in increase in cost, complexity and legal implications while transferring personal data across borders. In recent customer engagements, we have come across following instances:

- Many Canadian provinces require personal data of "public bodies" to be stored within Canadian boundaries.
- Australia, Japan, and China require records be stored within their borders.
- Subject to limited exception, Argentina, Brazil, and Mexico restrict the disclosure of personal data outside the country without prior consent of the user.
- Client organizations are increasingly demanding very stringent local data storage requirements when submitting an RFP or negotiating contracts.
Almost all the data processing IT systems are facing the above issues while meeting business and legislative requirements. Cloud based solutions that work with PII data also require capabilities for supporting the rapidly evolving data storage and processing requirements. The combination of above mentioned regulatory, technical and compliance elements require robust and secure data hosting and processing solutions for the customer preferences.

Scope
The scope of this white paper is limited to outlining the cloud-based reference solutions for the geolocalization of Personally Identifiable data. For illustration purposes, Azure Cloud service references will be used. For the exact implementation, the solution requirements will be informed by other data requirements and policies (e.g., Privacy; Security; Archiving etc).

This Document
The document outlines requirements identified by a group of architects and few customer representatives at Coforge, their potential solution options. For delivering solutions with increased flexibility of data storage and/or security of data transit, operational cost implications is considered. The document also presents cloud-based solutions recommending most preferred solution subject to conditions.

Requirements Overview
Local Data Residency of Personally Identifiable Information (PII)
For any given customer, it must be possible to physically store all PII data in a geographical area, whether that is a specific continent, country, or region within a country, and whether this storage is legally mandated (Data Localization law), legally encouraged (Data Sovereignty laws), or simply the choice of the Client (Data Residency).

Subject to customer specific requirements, following type of data meets the definition of personal data:
- User PII (First and last name, phone number, email address etc.)
- Any free text field that is stored against a candidate record
- Any file that relates to the candidate, e.g., Resume/CV
- Any audit record that contains User PII, including a users’ IP address

Personal data definition: Personal data is any information that relates to an identified or identifiable living individual. Different pieces of information, when collected leads to the identification of a particular person, also constitutes as personal data.

Privacy laws around the globe can vary in their definition of PII/Personal Data. For example, GDPR defines personal data as any information relating to an identified or identifiable natural person. Whereas the California Consumer Privacy Act (CCPA) defines personal information as “information that identifies, relates to, describes, is reasonably capable of being associated with, or could reasonably be linked, directly or indirectly, with a particular consumer or household.” In addition to data that is more intuitive (name, email address, social security number), PII under CCPA includes inferences drawn to create a profile about a consumer reflecting the consumer’s preferences, characteristics, psychological trends, predispositions, behaviour, attitudes, intelligence, abilities, and aptitudes.

In Canada, Personal Information Protection and Electronic Documents Act (PIPEDA)’s definition includes any factual or subjective information, recorded or not, about an identifiable individual. This includes opinions, evaluations, records of a dispute between a consumer and a merchant, and intentions (e.g., intentions to acquire goods or services, or change jobs).
China’s PIPL’s definition of PII follows that of the GDPR, though for sensitive data and is much broader including "information that once leaked or abused may cause damage to personal reputation or seriously endanger personal and property safety" as well as race, nationality, religion, biometric information, health, financial account, personal whereabouts, and other information.

These differences in definition can create challenges in development of a globally applicable policy on personal data localization; therefore, the solution must provision for geo-specific adaptation.

The Solution Options
The traditional architecture approaches while designing a cloud based or on-premises system, are centralized services, applications, and databases; whereas the new privacy and compliance requirements need architects to think otherwise. The balance between centralized, cost-efficient solution and a decentralized privacy compliant solution needs to be achieved by bearing in mind the nature of system and budget.

The solution largely depends on the amount of engineering effort budgeted for the initiative and current state of the system. There would be more options for an IT system that is being built / designed compared to an existing system that is being tuned to accommodate data localization requirements.

Following are few architecture options that have been explored and applied by our internal practice teams; all these solutions are relevant in their own space, subject to customer, project, and other circumstances.

Option 1: A central system, with geo-spread DB.
Solution Overview
All services are hosted in a central environment. The central environment acts as the ‘default’ environment and other environments are created when required for serving a new region.

- The applications and services are hosted in a central location.
- The databases are localized to all the regions where data residency requirements exist.
- The services and applications are connected to the regionally located databases based on the user’s location.
- The analytics data can still be centralized if anonymized data is stored in a central Datawarehouse or data mart etc.
- The recently fetched data from databases is cached to reduce frequency of cross region data transfer.
- All (non-PII) master data is cached alongside the services in a central region.
- Any updates to cached data is routed to the regional data stores, using a synch mechanism.
- All operational reports are generated locally, consolidated reports are generated from a central warehouse.

Benefits
- The centralization of services and applications help to monitor the system centrally.
- Streamlined support process due to centralization and application environment needs to be maintained.
- Simplified features roll out and upgrade of services, simplified DevOps.
- Licensing and subscription costs can be optimized by utilizing cloud elasticity.

Constraints
- Potential performance impact due to data storage and processing being distributed across regions
- Data exchange across regions that have regional firewalls blocking data transfer and, will need to be addressed separately.
- The locally cached data must be synchronized frequently for delivering services in real time.
- Users’ base location will be required to be resolved; this will need to be done using IP or user id.
- Non-functional aspects like audit, security logging must be stored regionally since these logs contain PII data. This adds to system complexity and therefore impacts maintainability.
- Recommended scenarios
- Line of Business applications where working dataset are not required to be fetched frequently – analytics apps.
- The services and applications can afford eventual consistency since data updates will need to be queued for performance reasons.
- Systems that are not custodians of PII data use it only for references – like training systems, recruitment systems.
- The user groups can be partitioned based on region.
Option 2: De-centralized system with locally hosted services and Data

Solution Overview
All services and all data are hosted and processed locally. The central environment will act as the 'default' environment and other environments will be created when required for serving a region.
- There is regional deployment of all services and databases and other repositories.
- Automatic redirection of a user will be done based on his / her source IP address.
- There would be a landing page for redirecting users in case they are logging from non-home region / geography.
- Master data replication will have to be designed separately; but this does not impact transactional performance.
- Consolidated data masked and replicated to central data mart / data lake for reporting.
- While viewing reports in central environment, Region specific redirection will have to be done for enabling operational reporting.
- For releasing new features, Applications and services will need to be deployed in multiple target environments to keep all regions current.

Benefits
- Having independent environments ensure full compliance since region specific tools and policies can be utilized.
- Simple architecture without any specialized engineering, resulting in low-cost operational support.
- Regional environments can be scaled up, down or decommissioned quickly, providing agility for services delivery to new regions.
- No performance penalties since data transfer across regions is not required.
Constraints
- Replicating the entire environment locally will incur higher technical architecture costs than other options.
- New environments will have to be created only when a new customer is onboarded, requiring ‘just in time’ investment
- The capacity and technical services plans can be selected and scaled up or down based on the expected volume of new business
- A centralised solution with multiple tenants will have to be scaled up
- A more complex system (options 1 or 3) would cost much more to develop and maintain, almost certainly more than replicating new localized environments on demand
- Creation of Disaster Recovery environment will provide more clarity about cost and support implications
- With multi region decentralized systems, operations will be complicated and upkeep of services across multiple environments will cause operational overheads.

Recommended scenarios
- Existing systems that cannot be subjected to re-engineering
- The compliance and data security requirements are ongoing and changing and need regional treatment.
- The user groups can be partitioned based on region; and regions can be allocated to the partitioned user base.
- Systems that need to process large chunks of data that includes PII and non PII Data.
- Systems requiring real time updates.
Option: A central system with PII data geo-located and rest of the data is central

Solution Overview
All services and non-PII data will be hosted centrally. PII data and repositories will be hosted regionally. The services will fetch PII data on demand and will cache the reference data.
- The PII data will be used only for references and if any update is required, it will be done asynchronously.
- All master, reference and transactional data will be centralized.

Benefits
- Having a central system hosting services and applications will simplify operations.
- The transactional data will be stored closer to the services. This helps deliver performant services.

Constraints
- The applications and services will need to switch data connections based on regions, and data (PI and Non PII). This complicates the system architecture and design.
- The end-to-end information lifecycle management will be complicated since the data is partitioned across regions.
Recommended scenarios
All services and non-PII data will be hosted centrally. PII data and repositories will be hosted regionally. The services will fetch PII data on demand and will cache the reference data.

- Systems that do not need to operate on PII data; and only need to refer to it.
- Heavy transaction data processing is required – this model helps by placing data closer to services.

Summary and Recommendations
While designing modern digital services delivery platforms, compliance requirements like GDPR and PII must be considered to ensure longevity of the solution. The customer and legislative demands are increasingly encouraging privacy by design principles. Modern architecture patterns like microservices are already compatible with above mentioned approaches since they readily allow data partitioning and separation of data processing from data storage.
Coforge is a leading global IT solutions and services organization which believes that real transformation cannot be driven by thinking in technology terms alone. With a mission to “Transform at the Intersect” it aims to bring both deep domain and deep emerging technologies expertise to achieve real-world business impact. A focus on very select industries, a detailed understanding of the underlying processes of those industries and partnerships with leading platforms provides us a distinct vantage. We leverage AI, Cloud and Insight driven technologies, allied with our industry expertise, to transform client businesses into intelligent, high growth enterprises.

Learn more about Coforge at: www.coforge.com