ACCESSING MORE DATA THROUGH TRUSTED EXECUTION ENVIRONMENT TO GENERATE NEW INSIGHTS

IMDA PET SANDBOX – HEALTHCARE SERVICES CASE STUDY

INFOCOMM
MEDIA
DEVELOPMENT
AUTHORITY

Contents

Business Case

1. The use case owner ("the Company") is a regional distributor of pharmaceutical products from pharmaceutical companies to storefronts (e.g. clinics and pharmacies). Its data analytics services provide insights to pharmaceutical companies about the movement of their products from manufacturing plants to hospitals, clinics and pharmacies. However, its data models do not have visibility of data about the "last mile" of such products' journey, i.e. their usage and treatment options.

2. **The Company has ecosystem data partners** (e.g. third-party administrators[1] (TPA), healthcare providers, digital therapeutics) across 13 countries in Asia. These data partners work with the Company to make its data models more precise and complete and accurate for the following analysis, e.g.
   - Rate of growth in patients diagnosed with specific conditions (e.g. Diabetes & CKD) based on the sales of the type of drugs (e.g. SGLT2i) associated with the conditions.
   - Growth of pharmaceutical brands among new patients vis-à-vis existing patients.
   - Trend of patients switching to competitor pharmaceutical brands.

3. However, the Company's data partners **commonly faced challenges sharing datasets in their original form** because of a few barriers:
   i. **Data protection regulations** in jurisdictions which do not allow transfer of personal data without consent and obtaining consent from the individuals concerned is not considered practical.
   ii. **Competition laws** could be infringed in the disclosure of product distribution and transaction data within the pharmaceutical ecosystem.

4. A manual workaround those challenges would be to deploy sales forces in the field to gather data by sampling usage and switching patterns. However, such efforts would be costly and collect outdated and less accurate 'last mile' data. Therefore, the desired outcome of using a PET-based solution would be to digitally onboard ecosystem data partners, build joint insights, while staying compliant to applicable regulations.

Solution Architecture

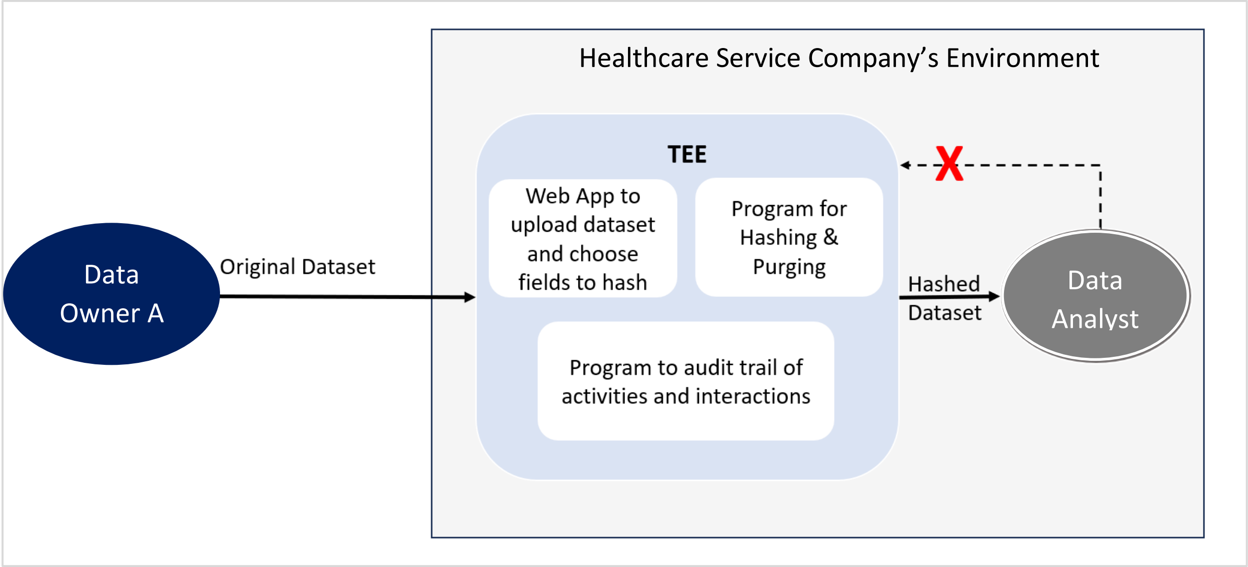
5. **'Confidential Computing', also known as 'Trusted Execution Environment' (TEE)** in the PET industry, was considered for this proof of concept (POC). TEE is a type of PET where isolated environments or enclaves guarantee that the data and applications inside them

[1] TPAs provide administrative services for health plans of clinics or medical benefits of corporates.

remain protected, and only the results of the applications run on the data can be exported. TEE also allows user access and query types to be limited programmatically.

6. In the POC, a TEE-based solution was designed for the use case owner's ecosystem partners (referred to as "Data Owner A" below) to interact in the following 3 steps (see Figure 2).
   i. Data Owner A transfers data in its original form, which may include personal data (PD), into the TEE. This TEE resides in the Company's environment.
   ii. When this original data is in the TEE, the data remains inaccessible in any form to the Company or any other 3rd party. Data Owner A hashes the fields containing personal data (and any other sensitive fields) using the SHA256 hashing algorithm.
   iii. The hashed fields and the remaining non sensitive and non-PD data are transferred to a file space outside of the TEE and the copy of the original data within the TEE is deleted.

Figure 2 – TEE-based solution architecture


Healthcare Service Company's Environment
TEE
Web App to upload dataset and choose fields to hash
Program for Hashing & Purging
Program to audit trail of activities and interactions
Data Owner A
Original Dataset
Hashed Dataset
Data Analyst
X

7. **The POC included 3 safeguards** to ensure that the original data contained within it cannot be read, modified, or otherwise accessed in any form by the Company (as host of the TEE):

   i. **The TEE contains the necessary application for Data Owner A to transfer in its original data and transfer out the hashed dataset, while protecting the original data.**
      - The application, that helps Data Owner A to hash sensitive data fields and transfer the hashed dataset, is created as a docker image, which in turn is converted to an IntelSGX compatible docker image.
      - The conversion process inserts the 'EnclaveOS' operating system in the docker image, creating a Memory and an Encrypted File System (EFS). The Memory and the EFS form the foundation of the TEE which cannot be accessed by the Company

(or any other external party) regardless of how high a privilege one may have in the larger environment hosting the TEE. This protects the original data (unhashed form) and prevents data access by the Company.

ii. **An auto-destruct is programmed into the web application in the TEE which purges the instance of the original data** after a pre-set time, regardless of whether the hashing is complete.

- If the dataset file is left by Data Owner A as draft (before hashing) in the TEE, the web application checks for such files unattended for longer than a pre-set time period (configured at 1 minute for the POC) and permanently deletes them.
- If the dataset file is hashed by Data Owner A, the web application is programmed to permanently delete the original dataset file and transfer out the hashed dataset file to the Company's data analytics environment.

iii. **The owner of the TEE gets a read-only audit log of activities and interactions** in the TEE which it can share with Data Owner A as assurance. This is by means of a 'Confidential Computing Manager' (CCM).

- The CCM can whitelist the types of applications to be run in the TEE, which are agreed to between Data Owner A and the Company ahead of the data transfer and hashing.
- The Company and Data Owner A can also monitor the timestamps and sources of activities in the TEE by means of the CCM.
- Further, the CCM allows them to view nodes, attested by an attestation system built into IntelSGX (by 'Intel Attestation Service'), which interact with the contents of the TEE.
- This means that a malicious actor will not be able to gain access to the sensitive information inside the TEE unless the CCM issues an attested node issued by Intel's Attestation Service. Neither the Company nor Data Owner A have control over this attestation process.

Regulatory Considerations

8. Based on the design of the POC, the Company sought **guidance from the Singapore Personal Data Protection Commission** (PDPC) on conditions of compliance for both it and its ecosystem data partners, i.e. a third party organisation or "TPO".

9. PDPC's view was that the TPO is **engaging the Company as a data intermediary (DI) to provide hashing services through its TEE** and web application[2]. Express consent is not

[2] An analogy is an organisation engaging a cloud service provider as a data intermediary to provide cloud services.

necessary for an organisation to share personal data with its DI to process personal data on its behalf, provided that the personal data is not used by the DI for other purposes without the consent of the individual[3]. In this case, since the Company will not use or even access the personal data in the TEE for other purposes, the TPO may transfer personal data into its TEE without obtaining consent from the individuals.

10. PDPC provided three additional comments under its regulatory guidance to the Healthcare Services Company.

    i. **The contract between the TPO and Company should make clear what scope of work** the latter is to perform on the TPO's behalf and for its purposes, e.g. providing the TEE and the web application for hashing, and each party's responsibilities and liabilities in relation to the transferred personal data.

    ii. **As a DI of the TPO, the Company is subject to the Protection, Retention Limitation, and Data Breach Notification Obligations** under Singapore's Personal Data Protection Act (PDPA). The technical safeguards put in place to protect the transferred data, and the web application which ensures the deletion of the transferred data (either when the hashed data is transferred out or upon the expiry of a period of time specified by the TPO), help the Company meet its Protection and Retention Limitation Obligations in the TEE implementation. Under the Data Breach Notification Obligation, where a data breach[4] is discovered by the Company, it would be required to notify the TPO without undue delay from the time it has credible grounds to believe that the data breach has occurred.

    iii. **The Company and the TPO may wish to ensure that the hashes generated should be reasonably strong** (e.g., by using industry-standard algorithms and incorporating a salt) to protect the data, particularly in the case of data that follows pre-determined formats or parameters such as NRIC numbers and race. While hashes are cryptographically generated strings that serve as irreversible one-to-one representations of the data that was hashed, proper safeguards need to be implemented to prevent attackers from identifying individuals through inferences from pre-computed tables.

[3] See PDPC's Guide to Data Sharing, at paragraph 1.8.

[4] "data breach", in relation to personal data, means —
(a) the unauthorised access, collection, use, disclosure, copying, modification or disposal of personal data; or
(b) the loss of any storage medium or device on which personal data is stored in circumstances where the unauthorised access, collection, use, disclosure, copying, modification or disposal of the personal data is likely to occur.

Feasibility Assessment

11. **The Company managed to access more data from data partners and create new data models to derive insights.** With the POC, the Company was able to onboard 2 data source partners which contributed data on diagnoses and prescriptions to the order of 20 GB. The typical time to educate and onboard data partners was about 3 months. With access to the hashed dataset, the Company's analysts were able to build 5 data models on the use of pharmaceutical drugs, trends in medical treatments and brand switching behaviour of patients.

12. **There were 2 key technical challenges in the POC.**

    i. **Building applications for data partners within TEE:** It was imperative for the Company's data partners, who had little technical expertise in hashing the data and using the TEE, to be provided with functions and features to ease the use of TEE. However, the Company faced challenges building applications compatible with the TEE. Running scripts in common libraries were not directly compatible with EnclaveOS. The applications needed to be built using IntelSGX software development kit to resolve this incompatibility. In this pilot's case, the TEE vendor had to actively help the Company overcome this challenge.

    ii. **Limited Capacity of TEE:** The POC demonstrated that the TEE had limited capacity to support multiple applications due to the nature of the operating system and the memory it carves out for the enclave in the wider computing environment (Azure SGX in the case of this POC). For supporting the web application mentioned earlier, the TEE required running a virtual machine of 64GB at a cost of about USD 11,000/year, as compared to USD 3,000/year for a similar application on commercial cloud.

Conclusions and Next Steps

13. Based on its experience developing the POC, the Company concluded that while confidential computing significantly widened its access to useful data sources and ability to create "last mile" data insights, the costs and resources borne by it to develop compatible applications in the TEE were significant.

14. Educating data partners about how the TEE-based solution protects data, and the governance measures that the host of the TEE (i.e. the Company in the case of this POC) needs to take as the DI add some complexity to adopting the solution at scale. This complexity is expected to slow down the rate of adoption of TEEs but could improve as more awareness about TEEs and its use cases spreads in the healthcare ecosystem.