ENABLING DATA AVAILABILITY THROUGH AUTOMATED DATA MANAGEMENT

IMDA PET SANDBOX – GRAB CASE STUDY

INFOCOMM
MEDIA
DEVELOPMENT
AUTHORITY

Contents

Business Use Case 2

Objective of POC 2

Methodology 3

Overview of the solution 3

Technical Learnings 7

Regulatory Learnings 9

Conclusions and Next Steps 12

Business Use Case

1. As a digital company with regional operations, GrabTaxi Pte. Ltd. (“Grab”) collects and owns data from its complex operations spanning across multiple domains, including personal data. Before these data could be used for internal analysis, the data underwent a manual process of tagging each dataset and classifying into “Restricted” (i.e. containing personal data) or “Unrestricted” (i.e. no personal data). As part of data management practices, users had to seek permission from the data owners to access any data within “Restricted” datasets. These approvals impacted the productivity of owners and requestors and was a key obstacle impeding time to market and data-driven decision making.

2. With more than 200K unique data column names in its data lake alone, on top of its growth in data volume, Grab faced certain challenges to its existing manual method of tagging, classification and access to personal data including:

   a. High overheads related to tagging and access approvals;

   b. Misinterpretation of personal data-related rules, resulting in data misclassified as “Restricted”, which could have been otherwise made available for further use.

3. Grab explored both rule-based tools and LLM-based PET tool to automate the tagging and classification of data. While rule-based tools could detect personal data, data was only tagged on table level and was insufficient to meet Grab’s requirements for column-level tagging. In addition, changes in personal data parameters (e.g. due to regulatory updates) usually required additional engineering efforts (typically involving Machine Learning engineers) to finetune the solution. Comparatively, LLM-based PET tool enabled granular tagging at column level and the natural linguistic capabilities of LLMs made it easier for data governance officers to centrally update data definitions based on various regulations across regional offices.

Objective of POC

4. Grab developed an in-house PET solution that comprises LLM-based metadata tagging engine and data protection techniques to automate column-level metadata tagging and data classification. With column-level tagging and data classification, personal data within “Restricted” datasets are protected and users can access the non-sensitive data immediately for quick insights, instead of waiting for data owner to grant access to these datasets.

Methodology

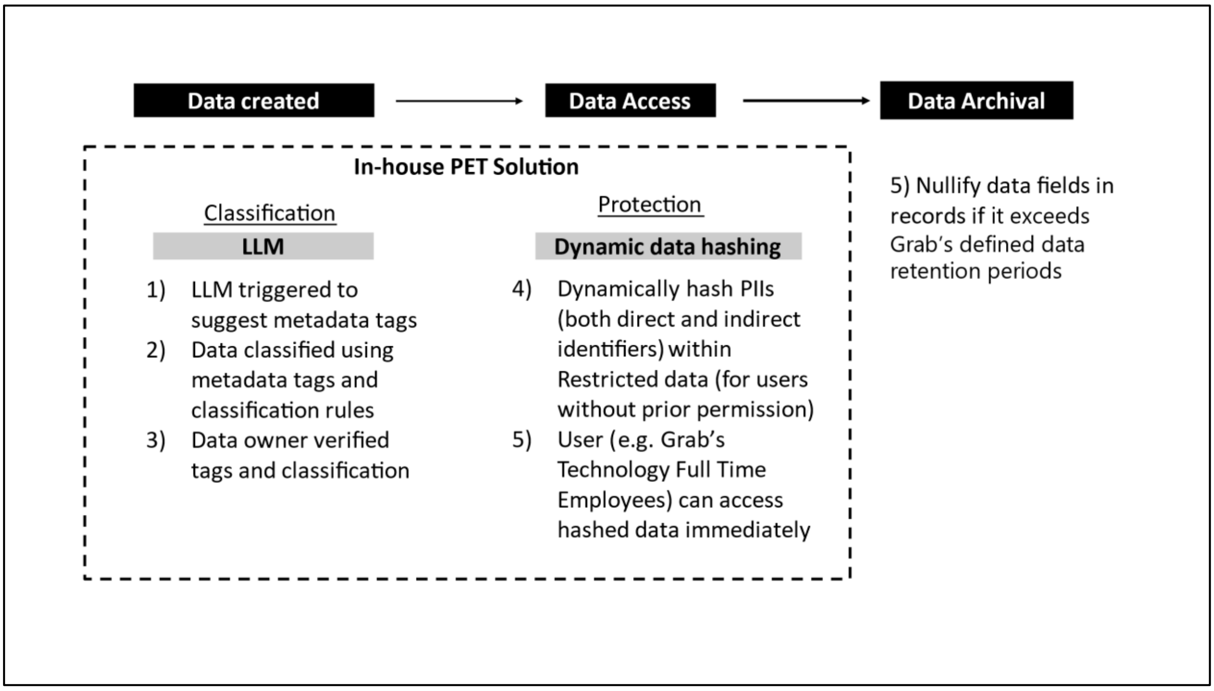
5. For this POC, Grab implemented the PET solution using 2-4 AWS pods which can process up to 5,000 tables/ day that can sufficiently meet their data generation speed and hosted the solution on premise to prevent sensitive data from being sent to LLM for training purposes.

6. The solution is tested on ~100K data in the data lake, where the LLM-based metadata tagging engine would suggest tags for columns containing Personally-Identifiable Information (PII). Based on these tags, data would be classified as “Restricted” (if it contains any PII (direct identifier) , or at least three (3) PII (indirect identifiers)) and the others will be classified as “Unrestricted”.

7. Data protection technique is applied to “Restricted” data within a data lake to enable differentiated access, depending on users’ access permissions. For users without prior permission to access “Restricted” data, PIIs would be hashed, allowing immediate access to the non-PII columns in the data lake.

8. The measurement of success for Grab is an increase in the amount of data available in the data lake for internal use without requiring additional approvals.

Overview of the solution



1) LLM triggered to suggest metadata tags
2) Data classified using metadata tags and classification rules
3) Data owner verified tags and classification
4) Dynamically hash PIIs (both direct and indirect identifiers) within Restricted data (for users without prior permission)
5) User (e.g. Grab’s Technology Full Time Employees) can access hashed data immediately

5) Nullify data fields in records if it exceeds Grab’s defined data retention periods

Figure 1: Solution Overview

The solution has 2 components:

LLM-based classification engine

9. **Generate tags:** When data is created in the data lake, the classification engine will read the column header and suggest the most relevant metadata tags from a list of pre-defined tags for columns containing PII. For non-PII columns, a default [None] tag is used.

Type of PII	E.g. Metadata tags	Column Name
Direct Identifier	[Personal.ID]	E.g. "NRIC", “FIN”, “License Plate”
	[Personal.Name]	E.g. “Username”
	[Personal.contact_info]	E.g. “email”, “phone”,”address”
Indirect Identifier	[Geo.Geohash]	E.g. "latitude”
	[Personal.Traits]	E.g. “age”, “gender”
Non-PII	[None]	E.g. If none of the above tags can be assigned

Table 1: How types of PII map to metadata tags

10. **Generate classification:** Data will be classified as “Restricted” or “Unrestricted” based on the following classification rule:
“Restricted” if contains any PII (Direct Identifier) or at least 3 PII (Indirect Identifier] ; else “Unrestricted”

Example of how classification will work :

Dataset 1 : Contains PII (Direct Identifier]
Classification : “Restricted”

Column Name	NRIC	Age	Transport Type
Metadata tag	[Personal.ID]	[Personal.Traits]	[None]
PII type	*Direct Identifier*	*Indirect Identifier*	*Non-PII*
E.g.	S1234567A	45	Taxi

Dataset 2: Contains less than 3 PII (Indirect Identifiers]
Classification : “Unrestricted”

Column Name	Gender	Age	Transport Type
Metadata tag	[Personal.Traits]	[Personal.Traits]	[None]
PII type	*Indirect Identifier*	*Indirect Identifier*	*Non-PII*
E.g.	Male	45	Taxi

11. **Verification of tag and classification** by data owners before the metadata is made available in the data lake.

<u>**Dynamic hashing to protect PII**</u>

12. **Dynamic hashing** After the data owners verify the data classification, the data can be accessed by users in the data lake. When users want to access “Restricted” data in the data lake, an API call is done to check the user’s permissions for this data. If they have no permission to access “Restricted” data, both PII (direct and indirect identifiers) are hashed using SHA256 and user can access hashed data automatically. For “Unrestricted” data, no hashing will be applied.

13. SHA-256 protects PII by cryptographically transforming the data into irreversible values. As it is computationally infeasible to reverse the process and retrieve the original input from the hash value, the PII remains protected even when users access “Restricted” data with hashed columns.

Example of how hashing protects the PII
Dataset : Contains PII (Direct Identifier) and PII (Indirect Identifier)
Classification : “Restricted”

Before

Column Name	**NRIC**	**Age**	**Transport Type**
Metadata tag	[Personal.ID]	[Personal.Traits]	[None]
PII type	*Direct Identifier*	*Indirect Identifier*	*Non-PII*
E.g.	S1234567A	45	Taxi

After

Column Name	**NRIC**	**Age**	**Transport Type**
Metadata tag	[Personal.ID]	[Personal.Traits]	[None]
PII type	*Direct Identifier*	*Indirect Identifier*	*Non-PII*

E.g.	70F2B95BDB288B37DE66EF0548F97F124E84D68F700EFDD893F7AA48462A	420002158111BFF8ADB3347D84029E480F45E60E1869B454B6ADAC3345F	Taxi

14. **Nullification:** For personal data that has been stored/ archived in the data lake, periodic review is conducted to delete any PII within the records that has reached its retention period. For each individual record that has reached its retention period, all PII is permanently deleted from the record using an in-house application. This further protects PII from being retained in perpetuity beyond business or legal needs.

Before

Column Name	**NRIC**	**Age**	**Transport Type**	**Year of Generation**
Metadata tag	[Personal.ID]	[Personal.Traits]	[None]	[None]
PII type	*Direct Identifier*	*Indirect Identifier*	*Non-PII*	*Non-PII*
E.g.	S1234567A	45	Taxi	2010
E.g.	F7654321Z	21	Car	2020

After

Column name	NRIC	Age	Transport type	Year of generation
Metadata tag	[Personal.ID]	[Personal.Traits]	[None]	[None]
PII type	*Direct Identifier*	*Indirect Identifier*	*Non-PII*	*Non-PII*
E.g.	<null>	<null>	Taxi	2010
E.g.	F7654321Z	21	Car	2020

Technical Learnings

LLM related

15. Hallucination is a phenomenon where the model generates text that is incorrect, nonsensical or not real but presented as facts. In Grab's context of metadata tagging, LLM could potentially deviate from the list of predefined tags or format when generating its outputs. In order to reduce hallucination, LLM is configured to only suggest pre-determined tags for columns containing PII and [None] tag for data columns that do not contain PII. If the in-house application detects tags not within pre-determined tags, LLM would reattempt to generate the tags. Subsequently, during the tag verification stage, the data owners will check and correct the tags that are not within the pre-determined tags.

16. Tokens, or pieces of text, are used by LLM to understand the input text and generate responses, however there is limitation to the total capacity of the LLM to process input and generate output. This limitation is referred to as token limitations. In Grab's context, this means that larger tables cannot be processed within a single session as the input alone would exceed the total token limitation. To work around the challenge of token limitation, Grab added in an additional step to first determine if batch processing is required, and by processing 40 columns in each batch, LLM was able to successfully process tables with up to 5000 columns.

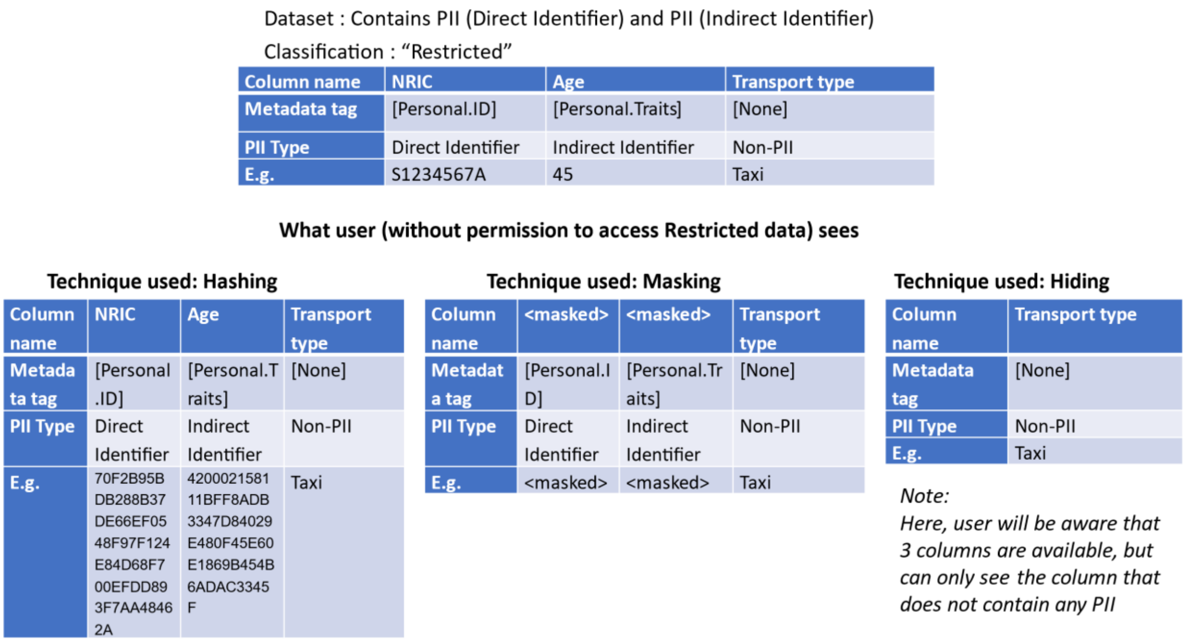
17. However, due to the "batch" processing of the table, LLM does not 'remember' that the table is batch processed and processes each batch as a 'new' table, where LLM may suggest inaccurate tags when columns with relationships are processed in different batches. This challenge is overcome by including table name & small sample of rows from the dataset to provide context to LLM when processing each batch, hence reducing the probability of generating inaccurate tags during batch processing.

18. For example, in processing the column name "date", without context, the metadata tag will likely be [None]. if LLM considers the table name "credit.card.info", it will likely generate the metadata tag [Financial.Info]. The LLM could also infer based on the sample rows format of MM/YY, that this refers to expiry dates on credit cards.

Protection related

19. Grab explored various methodologies to "hide" PII within "Restricted" data : masking where the user would only see the column name with masked values and "hiding" the entire column with PII such that the data user only has knowledge that other columns exist within the data but does not know the column names. Hashing was selected as it

offered Grab the highest data utility as the deterministic output maintains record level inference (e.g. to allow building user profile) while ensuring that no PII is revealed or used.



Dataset : Contains PII (Direct Identifier) and PII (Indirect Identifier)

Classification : “Restricted”

Column name	NRIC	Age	Transport type
Metadata tag	[Personal.ID]	[Personal.Traits]	[None]
PII Type	Direct Identifier	Indirect Identifier	Non-PII
E.g.	S1234567A	45	Taxi

What user (without permission to access Restricted data) sees

Technique used: Hashing

Column name	NRIC	Age	Transport type
Metadata tag	[Personal.ID]	[Personal.Traits]	[None]
PII Type	Direct Identifier	Indirect Identifier	Non-PII
E.g.	70F2B95BDB288B37DE66EF0548F97F124E84D68F700EFDD893F7AA48462A	420002158111BFF8ADB3347D84029E480F45E60E1869B454B6ADAC3345F	Taxi

Technique used: Masking

Column name	<masked>	<masked>	Transport type
Metadata tag	[Personal.ID]	[Personal.Traits]	[None]
PII Type	Direct Identifier	Indirect Identifier	Non-PII
E.g.	<masked>	<masked>	Taxi

Technique used: Hiding

Column name	Transport type
Metadata tag	[None]
PII Type	Non-PII
E.g.	Taxi

Note:
Here, user will be aware that 3 columns are available, but can only see the column that does not contain any PII

Figure 2 : Examples of outputs from different methodologies explored

Other governance measures in deploying PET solution

20. As LLM is inherently a machine learning model, Grab has implemented other AI governance measures for the responsible use of AI in classifying data, such as:

<u>Establishing checks and balances</u>

a. Approvals required for data owners to overwrite LLM-generated tags & classification, especially where data is classified from “Restricted” to “Unrestricted” to ensure that there is no inadvertent access to any PII that might exist in the dataset.

b. Monitor LLM’s performance through rate of LLM tag verification and dashboard that informs of LLM’s accuracy by measuring frequency of data owner overwriting LLM’s suggestions. A hard coded business rules engine also periodically scans the data to ensure all PII has been tagged. Data Governance Office also performs regular sampling checks to assess the LLM performance.

c. Monitor user activity within the data lake (e.g. sustained querying of specific dataset types) through periodic reviews of dashboard and logs by the Data Governance Office.

Human in-the-loop for AI-augmented decision-making

d. Prompt and incentivise data owners to verify data classification tags in a timely manner. While metadata tag and classification generation are automated, verifying metadata tags and data classification in a timely manner remains a key action for data owners who are ultimately accountable for the accuracy of their metadata.

Limit data for model development to mitigate unintended output

e. Internal policy to disallow LLMs to self-learn from revision of tags as the tag revision may not necessarily adhere to internal data governance rules. For example, if LLM is allowed to self-learn from manual tag revisions that involve removing a PII related tag (e.g. related to email), LLM might stop predicting a PII tag for columns that contain such PII, which deviates from the current data governance definitions.

Regulatory Learnings

21. Grab sought Practical Guidance (Guidance) from the Personal Data Protection Commission (PDPC) on the following:

a. Whether data records that contain hashed PII fields constitute personal data under the PDPA.

b. Whether LLM can be relied upon to do data classification.

c. Whether there are applicable exceptions to consent that Grab may rely on for the use and/or disclosure of customer personal data for data analytics and generation of business insights;

Whether the hashed data and nullified data constitute anonymised data

22. PDPC notes that Grab has implemented good data protection practices that combined technical and process controls. In particular, PDPC recognises that anonymisation techniques had been applied to enhance protection of personal data, while enabling the use of data for insights and data innovation (e.g., data analytics, data modelling). PDPC treats anonymisation as a risk-based process which includes applying both anonymisation techniques and safeguards (i.e., technical, process, administrative) to prevent re-identification. In determining whether personal data is anonymised, PDPC will take into account the following :

a. Whether all direct identifiers have been removed;

b. Whether indirect identifiers that can be used to re-identify individuals when matched with publicly available or proprietary information that the data recipient has access to have been altered or removed;

c. Whether there are additional safeguards implemented to restrict access and use of anonymised data to reduce the risks of re-identification (e.g., organisational structures, policies, processes); and

d. Whether there are periodic reviews conducted to assess adequacy of anonymisation techniques and risk management controls in relation to current state of technology, robustness of organisational, legal, processes and other non-technical measures to manage the risks of re-identification.

23. Based on the above, where both direct and indirect identifiers in a data record is nullified (i.e. removed), re-identification risk will be low. As such, PDPC would consider the data record to be anonymised.

24. For Restricted data, PDPC notes that hashing will be performed on both direct and indirect identifiers in each data record. While hashes are cryptographically generated strings that serve as irreversible one-to-one representations of the data that was hashed, proper safeguards should be implemented to prevent attackers from identifying individuals through inferences from pre-computed tables. Grab should ensure that the hashes generated are reasonably strong (e.g., by using industry-standard algorithms and incorporating a salt) to protect the data, particularly in the case of direct identifiers that follow pre-determined formats such as National IDs.

25. For Unrestricted data, PDPC notes that data records with fewer than 3 PII (indirect Identifiers) will not be hashed and can be accessed in the clear. While no direct identifiers are included in such data records, attackers/unauthorised parties may still re-identify individuals by querying and merging multiple records belonging to an individual and gaining access to data records with indirect identifiers in the clear. Such Unrestricted data would not be considered as anonymised. In particular, Grab will need to comply with the Protection Obligation by putting in place proper access controls and safeguards to protect such data, such as:

a. Monitoring of queries made, and/or random sampling/audit on persistent querying of data records with fewer than 3 PII (indirect identifiers);

b. Review of access policies (e.g., criteria for granting Restricted / Unrestricted user access rights, duration of access); and

c. Periodic review of user accounts to ensure that access policies are implemented (e.g., all the accounts are active and the rights assigned are in compliance with access policies, timely removal of user accounts when a user has left the organisation or update the user's rights when he/she has changed his/her role within the organisation).

Potential data protection risks arising from the use of LLM to classify data

26. PDPC recognises that data classification can be an effective tool to aid organisations in managing their data protection risks (e.g., by tailoring different sets of data protection measures/governance controls based on the data categories as defined by the organisation's internal classification policies).

27. In Grab's case, a combination of LLM tagging of data fields and rules-based classification is deployed to determine whether a data record qualifies as Restricted data. PDPC notes that there is a possibility that the LLM may not perform the tagging as intended, resulting in a downgrade in classification from Restricted to Unrestricted. This may increase the risk of "unauthorised user access" where a user gains access to supposedly Restricted data in the clear (when the PII within the data record should have been hashed). To address and to reduce the likelihood of inaccurate data tagging and classification, Grab has put in place safeguards to monitor the accuracy of LLM (e.g., periodically using hard coded business rules to counter check the tagging and classification of randomly selected data records), and to ensure that there are additional checks (e.g., manual verification) on the classification output.

Applicable exceptions to consent under the PDPA

28. Where relevant, Grab may consider relying on the following PDPA's exceptions to the Consent Obligation when using personal data:

    a. Business improvement exception is likely to apply where Grab's use of personal data is to generate insights to improve or develop new goods or services, or to better understand customer preferences and behaviour etc. To rely on the exception, Grab will need to ensure that the purpose cannot be reasonably achieved without using the personal data in an individually identifiable form, and that a reasonable person would consider the use of personal data for such purpose appropriate in the circumstances. Grab may also rely on the business improvement exception to share personal data, without consent, between entities belonging to a group of companies (e.g., Grab group).

b. Legitimate interests exception is likely to apply where Grab's use and/or disclosure of customers' personal data is for purposes such as fraud detection and preventing misuse of Grab's services. To rely on this exception, Grab will need to assess the adverse effect of the use and/or disclosure of personal data and ensure that the legitimate interests (i.e., benefits to Grab, other organisations, or wider segment of the public) in doing so outweigh any adverse effect on the individual

Conclusions and Next Steps

29. Grab validated the feasibility of using LLM to automate tagging and classification to increase data availability without the need for further access approvals. Overall, there is improved data availability within the data lake from ~50% ("Unrestricted" data) to ~85-90% ("Unrestricted" data, including non-PII within "Restricted" data) without needing to seek further access permissions.

30. In scaling to production, Grab also incorporated the following areas:

    a. Include tagging and classification of business sensitive data, where access is required to be managed in a similar manner to PII

    b. Due to Grab's real-time operational data needs, there is also a need to integrate with real time data streams using an additional code as there are typically no metadata tags associated with real time data streams. This allows the data streams to be processed and classified as the data are being generated. Involving other internal system owners ensures that LLM generated tags & classification are supported in data transmission across the entire data chain, which comprises several real time / online/ offline systems as well as backwards integration with legacy systems.

    c. Grab is also considering expanding the use of LLMs beyond tagging to include documentation (e.g. generating user-friendly descriptions table and columns) of other sensitive data related to financial and operating metrics. When the LLM development is sufficiently mature and accepted, there is potential to consider shifting the human review process to after the data is made available within the data lake as opposed to the current dependency of human review and verification.

31. Overall, the LLM has demonstrated that column-level tagging can be done efficiently and effectively. This forms a foundation for Grab to shift from the current role-based access into attribute-based access, which allows dynamic, granular authorisation decisions based

on a wider range of attributions such as changes in user attributions (e.g. data engineer, full time employee) or context (e.g. change in project needs), allowing Grab to make faster responses to market needs.