Data Vault model solution to Personally Identifiable Information

Scott Maner, Muhammad M Rana, Narayanan Veliyath, Vladan Jovanovic

Abstract - *The data vault model has remained relatively unchanged since the advent of Data Vault 2.0. While this model has remained viable through its lifetime, the necessity to further keep information safe has begun to prompt change. This paper will explore the practical benefits and general method of how to alter the data vault model to better fit the needs of today's consumers.*

Keywords: Data Vault, Encryption, Personally Identifiable Information (PII)

INTRODUCTION

In today's society, information is considered to be one of the most powerful tools one can have. However, simply acquiring and analyzing data is not sufficient. Like any other object of value, it needs a place to be reliably stored. The industry's solution to this is known as a data warehouse. A data warehouse is specifically designed to deal with the long-term storage of data, often from multiple sources. A specific model employed in this research is the Data Vault 2.0 model. This model differentiates itself from other models because it is an insert only architecture. In layman's terms, this simply means that data can be put into the vault, but cannot be deleted or altered once inside. More recently however, events have occurred that have necessitated additions to this design model approach.

The General Data Protection Regulation (GDPR) passed within the European Union (EU) is a directive that aims to return control of PII to the individual. PII can be any information that can potentially identify a person, such as a name, birthdate, or even an email address. The directive states that the individual should have ultimate control of their data in any system, and should have the right to delete, change, or restore it at their own will. However, this creates a fundamental conflict with the data vault's design, which does not permit deletions of any kind. So how can the model be updated to adhere to the new guidelines?

Naturally, this raises the question: Why should the US, or indeed any country or business outside of the EU, care about this? The GDPR has already been officially passed, and has become effective on May 25, 2018[1][3]. U.S. companies that handle the personal data of individuals located in the EU are closer to confronting this new data security and privacy regime. Even if the companies have no establishments in the EU, organizations that handle even small amounts of EU personal data may be surprised to find themselves subject to the GDPR, and will soon find the need to take steps to bring themselves into compliance. Perhaps the biggest incentive for companies to be compliant would the significant fines that could be incurred. Fines under the GDPR can vary significantly, with a maximum of the greater of either €20,000,000 or 4% of annual worldwide turnover, depending on the seriousness of the violation[2].

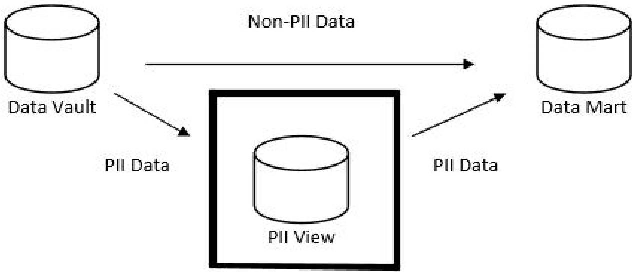
Related Works

To the best of our knowledge no other work regarding PII in data vault based data warehousing environment has been discussed.

Nevertheless, we consider current state of the practice, if not the state of the art, in databases regarding the PII protections is fairly represented by vendors such as Oracle's publicly available research into this field in response to the EU's GDPR[4]. Their approach is similar in that encryption is also a key aspect. However, Oracle has also strongly advocated for pseudonymization of the data. Furthermore, they are able to offer an extra layer of protection by storing master encryption keys within their Oracle Key Vault Service. It should be noted that Oracle has not publicly released any of the details of their efforts, and that their work is more focused in the field of databases rather than data vaults. While the Oracle Key Vault Service could certainly be used in our solution, the practical implementation details are considered to be beyond the scope of this paper

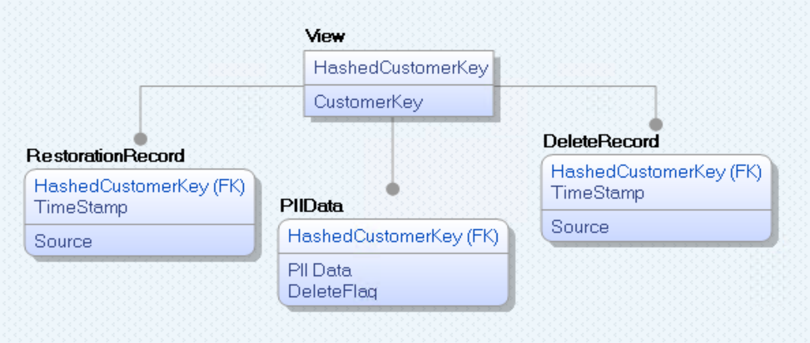
Research Methodology

The solution proposed is far more straightforward than one would initially imagine. Data vaults are made up of three entity types: Hubs, which contain surrogate keys as well as business keys; Links, which represent transactions or interactions between business keys and satellites, which contain the attributes[5][6]. In the current data vault model, these satellites would be where the PII are stored. In most cases before the GDPR, the PII stored would have minimal-to-no levels of protection. Under the GDPR, this would be a considerable offense. However, instead of reconstructing the entire data vault model, the more practical solution would be as simple as encrypting the data, as well as using a materialized table to access the PII.

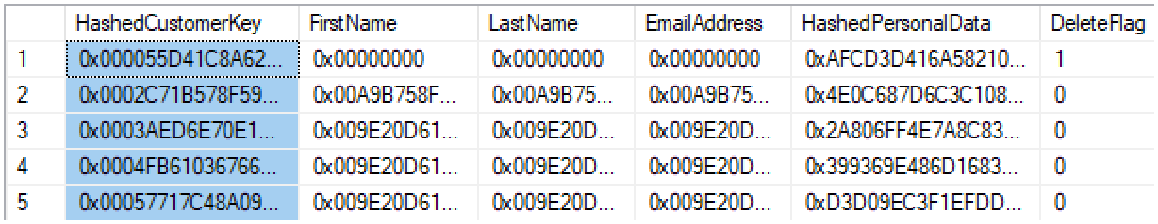


This would enable the data vault to safely hold PII, without compromising the fundamental nature of the model. All data requests would be handled through a virtual table located between the data vault and data mart. This materialized PII_view would allow access to the PII along with other generic information, without also giving access to the encrypted data stored in the vault. This offers the benefit of completing the request without having to disturb the contents stored in the vault. Additionally, the PII remains encrypted until it reaches the data mart. This means that any information intercepted while in transit or while in storage is unable to be decrypted by entities lacking the proper authorization.

As previously mentioned, the fundamental tenet of the data vault model is to allow only insertions. Other operations such as deletions or updates are not even considered under normal circumstances. However, under the GDPR, companies could be forced to delete PII stored in a data vault if so requested by their customer. The proposed model remedies this issue of requiring deletions with the introduction of a new group of entities between the data warehouse and data marts. These entities will act as a 'view' of all the PII data contained in the data vault. No other data would be stored within. The term 'other data', refers to non-PII related data, as within our PII_view we also store information relating to our PII data. This additional information acts as a record of all deletions and restorations received for the PII contained in the data vault. Additionally, the PII_view will also contain a hashed value. This value will represent all the PII data concatenated together, then hashed for reference.



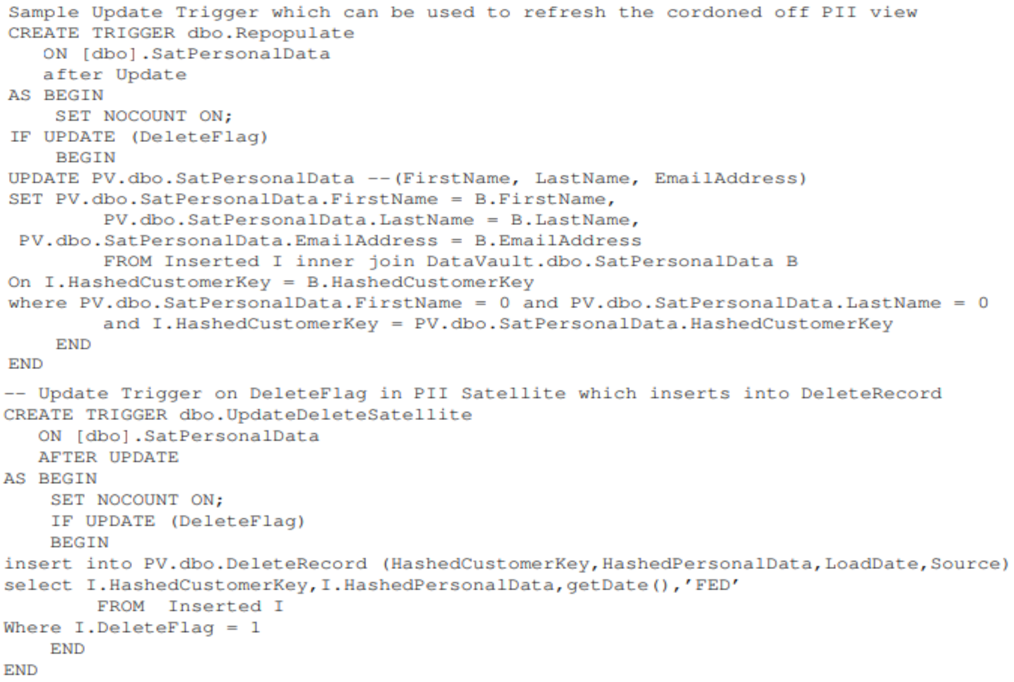
For the implementation of the proposed model, triggers were created within the PII_view to display the feasibility of our model. Triggers, which are essentially pieces of code that run when types of queries are run such as INSERTS, UPDATES, and DELETES, allow for a simple simulation of how this model would handle delete requests. Two triggers have been included: One which handles reconstructing the PII data from the data vault; and the other which acts to insert a record the deletion request. These triggers activate when the 'DeleteFlag' shown have been updated. For example when a delete request has come in from some source the database administrator would update the delete flag to 1 and models triggers would handle the rest. The results of such an operation on the PII_view can be seen below. One of the most important aspects of the proposed model is that the data mart no longer has any access to our PII data, and the data is untouched in the data vault. Through these triggers, the PII data can be deleted, updated or restored.



	HashedCustomerKey	FirstName	LastName	EmailAddress	HashedPersonalData	DeleteFlag
1	0x000055D41C8A62...	0x00000000	0x00000000	0x00000000	0xAFCD3D416A58210...	1
2	0x0002C71B578F59...	0x00A9B758F...	0x00A9B75...	0x00A9B75...	0x4E0C687D6C3C108...	0
3	0x0003AED6E70E1...	0x009E20D61...	0x009E20D...	0x009E20D...	0x2A806FF4E7A8C83...	0
4	0x0004FB61036766...	0x009E20D61...	0x009E20D...	0x009E20D...	0x399369E486D1683...	0
5	0x00057717C48A09...	0x009E20D61...	0x009E20D...	0x009E20D...	0xD3D09EC3F1EFDD...	0

The ability to restore PII data is one of the most fundamental features of this model. The person that owns the data may ask for the PII data at any point, even after it has been deleted.

Therefore, special care needs must be taken when handling such requests. In the proposed model, when PII data is ‘deleted’, it is not in fact removed from the vault. Instead, the model nullifies the PII data in the PII_view, and stores the hash of the nullified PII data in a satellite. Instead of keeping the deleted data, a hash of the deleted data is stored. This allows for the data vault to meet the consumer’s requests, while also allowing for future data restoration through the hash. For additional security, the hashes from the PII_view can be directly validated using the hashes in the data vault.



```
Sample Update Trigger which can be used to refresh the cordoned off PII view
CREATE TRIGGER dbo.Repopulate
   ON [dbo].SatPersonalData
   after Update
AS BEGIN
    SET NOCOUNT ON;
 IF UPDATE (DeleteFlag)
    BEGIN
UPDATE PV.dbo.SatPersonalData --(FirstName, LastName, EmailAddress)
SET PV.dbo.SatPersonalData.FirstName = B.FirstName,
        PV.dbo.SatPersonalData.LastName = B.LastName,
 PV.dbo.SatPersonalData.EmailAddress = B.EmailAddress
        FROM Inserted I inner join DataVault.dbo.SatPersonalData B
On I.HashedCustomerKey = B.HashedCustomerKey
where PV.dbo.SatPersonalData.FirstName = 0 and PV.dbo.SatPersonalData.LastName = 0
        and I.HashedCustomerKey = PV.dbo.SatPersonalData.HashedCustomerKey
    END
END
-- Update Trigger on DeleteFlag in PII Satellite which inserts into DeleteRecord
CREATE TRIGGER dbo.UpdateDeleteSatellite
   ON [dbo].SatPersonalData
   AFTER UPDATE
AS BEGIN
    SET NOCOUNT ON;
    IF UPDATE (DeleteFlag)
    BEGIN
insert into PV.dbo.DeleteRecord (HashedCustomerKey,HashedPersonalData,LoadDate,Source)
select I.HashedCustomerKey,I.HashedPersonalData,getDate(),'FED'
        FROM  Inserted I
Where I.DeleteFlag = 1
    END
END
```

References

1. The General Data Protection Regulation: A Primer for U.S.-Based Organizations That Handle EU Personal Data. (2018, March 07). Available at: https://wp.nyu.edu/compliance_enforcement/2017/12/11/the-general-data-protection-regulation-a-primer-for-u-s-based-organizations-that-handle-eu-personal-data/
2. Lazzarotti, J. J., & Costigan, M. (2018, January 08). Does the GDPR Apply to Your US-based Company? https://www.lexology.com/library/detail.aspx?g=3a02f14c-828b-47ba-bb91-cbddb41bbce3
3. En.wikipedia.org. (2018). *General Data Protection Regulation*. Available at: https://en.wikipedia.org/wiki/General_Data_Protection_Regulation
4. RAJASEKHARAN, D. (2017). *Accelerate Your Response to the EU General Data Protection Regulation (GDPR)*. [ebook] Available at: http://www.oracle.com/technetwork/database/security/wp-security-dbsec-gdpr-3073228.pdf
5. V.Jovanovic, I.Bojicic, C. Knowles, M.Pavlic “Persistent Staging Area Models for Data Warehouses” Issues in Information Systems V13, Issue 1(October 2012), pp 121-132
6. Jovanovic V., I. Bojicic “Conceptual Data Vault Model” Proceedings of the SAIS Conference, March 2012, Atlanta USA, pp. 131-136