

Deleting the Undeleteable

Behrooz Shafiee
Privacy Engineer



1.7 Million+ Merchants

~\$314 Billion Total Sales



Analytical Platform



Analytical Platform

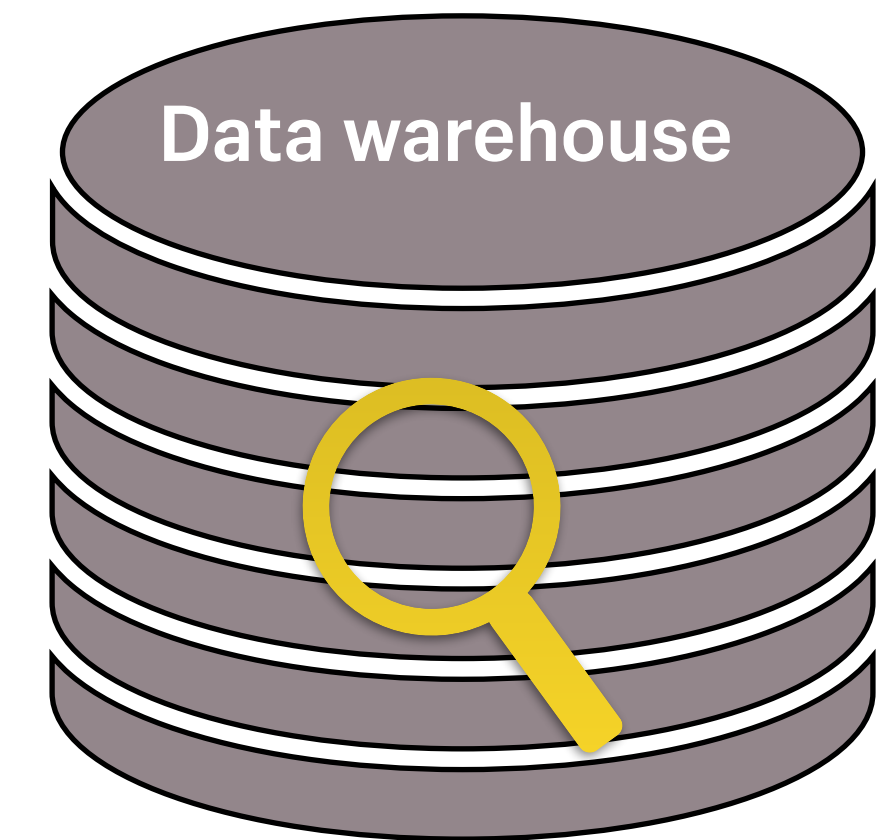


Analytical Platform



How to efficiently find/act on PII in data warehouse?

- **Immutable by design**
- **No guaranteed structure**
- **Missing privacy context**
- **Cascading failure chance**
- **Scale**



Context Collection

```
1  name: 3f_signup_request
2  doc: Triggered when users sign up to Shopify
3  version: 3
4  privacy_setting:
4    controller:
5      kind: SHOPIFY
6    subject:
7      kind: EMAIL
9  fields:
10 - name: email
11   type: string
12   doc: Email address for the merchant.
13   privacy:
14     type: EMAIL
15     tokenize: true
16 - name: remote_ip
17   type: string
18   doc: Request IP address.
19   privacy:
20     type: IP
```



Context Collection

- Structure



```
...
9  fields:
10 - name: email
12   type: string
13   doc: Email address for the merchant.
...
17 - name: remote_ip
18   type: string
19   doc: Request IP address.
...
```

Context Collection

- Structure
- Evolution



```
1  name: 3f_signup_request
2  doc: Triggered when users sign up to Shopify
3  version: 3
...
```

Context Collection

- Structure
- Evolution
- Privacy

```
...  
4  privacy_setting:  
4    controller:  
5      kind: SHOPIFY  
6    subject:  
7      kind: EMAIL  
9  fields:  
10 - name: email  
13   privacy:  
14     type: EMAIL  
15     tokenize: true  
16 - name: remote_ip  
19   privacy:  
20     type: IP
```



Pseudonymisation

- **Obfuscation (+enrichment)**

- IP (geo-locate)
- Useragent (parser)
- Coordinates
- Phone numbers
- Other identifiers

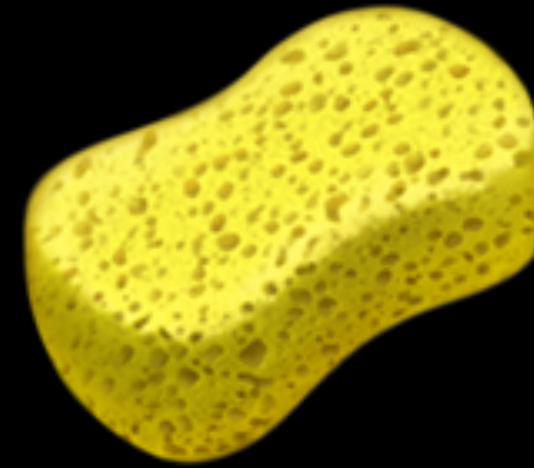
- **Tokenization**

- Exchange data with a **consistent** random token



```
{  
  ...  
  remote_ip: "22.33.44.55",  
  email: "behrooz@gmail.com",  
  ...  
}
```

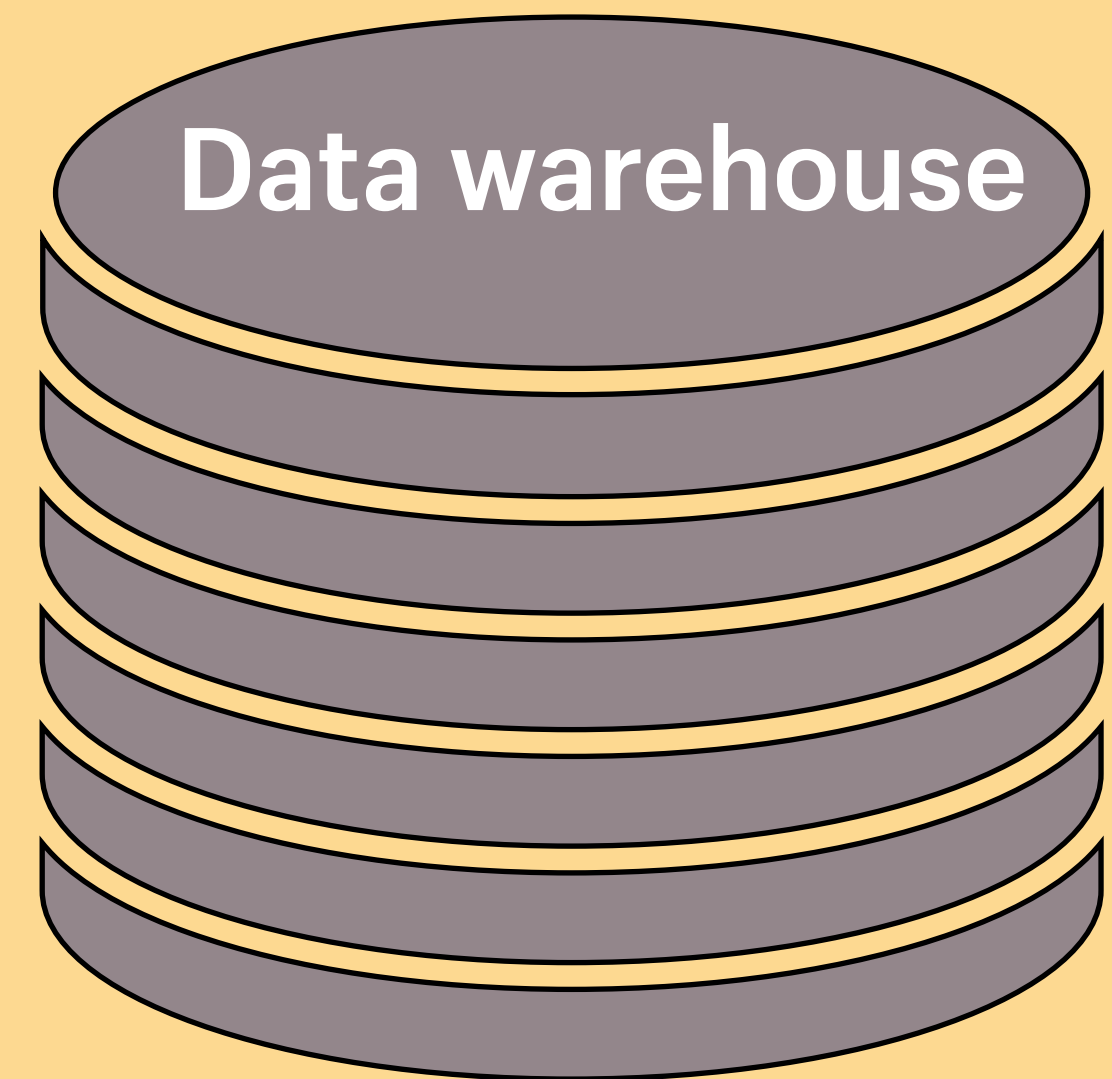
Scrubber



```
remote_ip: {  
  masked: "22.33.0.0",  
  geo_country: "Canada",  
  geo_city: "Ottawa",  
},  
email: {  
  obf: "redacted@gmail.com",  
  token: "43bdfs234sdf08ffpioaaxx",  
}
```

How does this help?

- **Non-personal data**
- **Obfuscated data**
- **Tokenized data**



How does this help?

- **Non-personal data**
- **Obfuscated data**
- **Tokenized data**



Tokenization
Vault



Email	Shop	Product	...
hooman@gmail.com	<i>allbirds</i>		
hooman@gmail.com			
hooman@gmail.com	<i>allbirds</i>		

allbirds Gymshark



Email	Shop	Product	...
hooman@gmail.com	<i>allbirds</i>		
hooman@gmail.com			
hooman@gmail.com	<i>allbirds</i>		

Tokenization
Vault



Data Subject	Controller	Token	PII

allbirds

Gymshark



Email	Shop	Product	...
	<i>allbirds</i>		
hooman@gmail.com			
hooman@gmail.com	<i>allbirds</i>		

Tokenization
Vault



Data Subject	Controller	Token	PII

allbirds

Gymshark



Email	Shop	Product	...
	<i>allbirds</i>		
hooman@gmail.com			
hooman@gmail.com	<i>allbirds</i>		

Tokenization
Vault



Data Subject	Controller	Token	PII
hooman@gmail.com	<i>allbirds</i>	token123	hooman@gmail.com

allbirds

Gymshark



Email	Shop	Product	...
token123	<i>allbirds</i>		
hooman@gmail.com			
hooman@gmail.com	<i>allbirds</i>		

Tokenization
Vault



Data Subject	Controller	Token	PII
hooman@gmail.com	<i>allbirds</i>	token123	hooman@gmail.com

allbirds

Gymshark



Email	Shop	Product	...
token123	<i>allbirds</i>		
			
hooman@gmail.com	<i>allbirds</i>		

Tokenization
Vault



Data Subject	Controller	Token	PII
hooman@gmail.com	<i>allbirds</i>	token123	hooman@gmail.com

allbirds

Gymshark



Email	Shop	Product	...
token123	<i>allbirds</i>		
			
hooman@gmail.com	<i>allbirds</i>		

Tokenization
Vault



Data Subject	Controller	Token	PII
hooman@gmail.com	<i>allbirds</i>	token123	hooman@gmail.com
hooman@gmail.com		token456	hooman@gmail.com

allbirds

Gymshark



Email	Shop	Product	...
token123	<i>allbirds</i>		
token456			
hooman@gmail.com	<i>allbirds</i>		

Tokenization
Vault



Data Subject	Controller	Token	PII
hooman@gmail.com	<i>allbirds</i>	token123	hooman@gmail.com
hooman@gmail.com		token456	hooman@gmail.com

allbirds

Gymshark



Email	Shop	Product	...
token123	<i>allbirds</i>		
token456			
	<i>allbirds</i>		

Tokenization
Vault



Data Subject	Controller	Token	PII
hooman@gmail.com	<i>allbirds</i>	token123	hooman@gmail.com
hooman@gmail.com		token456	hooman@gmail.com

allbirds

Gymshark



Email	Shop	Product	...
token123	<i>allbirds</i>		
token456			
token123	<i>allbirds</i>		

Tokenization
Vault



Data Subject	Controller	Token	PII
hooman@gmail.com	<i>allbirds</i>	token123	hooman@gmail.com
hooman@gmail.com		token456	hooman@gmail.com

allbirds

Gymshark



Multi-controller Tokenization*



Data Subject	Controller	PII	Token
hooman@gmail.com	<i>allbirds</i>	hooman@gmail.com	token123
hooman@gmail.com		hooman@gmail.com	token456
hooman@gmail.com		 222-333-4444	Token789
bob@gmail.com		IP 76.44.55.66	Token011

Multi-controller Tokenization*

Vault



Data Subject	Controller	PII	Token
hooman@gmail.com	<i>allbirds</i>	hooman@gmail.com	token123
hooman@gmail.com		hooman@gmail.com	token456
hooman@gmail.com		 222-333-4444	Token789
bob@gmail.com		IP 76.44.55.66	Token011

Multi-controller Tokenization*

Vault



Data Subject	Controller	PII	Token
hooman@gmail.com	<i>allbirds</i>	hooman@gmail.com	token123
hooman@gmail.com		hooman@gmail.com	token456
hooman@gmail.com		 222-333-4444	Token789
bob@gmail.com		IP 76.44.55.66	Token011

Multi-controller Tokenization*



Data Subject	Controller	PII	Token
hooman@gmail.com	<i>allbirds</i>	hooman@gmail.com	token123
hooman@gmail.com		hooman@gmail.com	token456
hooman@gmail.com		 222-333-4444	Token789
 bob@gmail.com		IP 76.44.55.66	Token011

Multi-controller Tokenization*



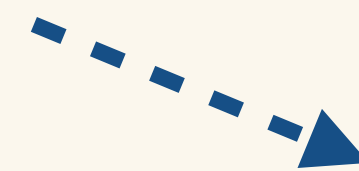
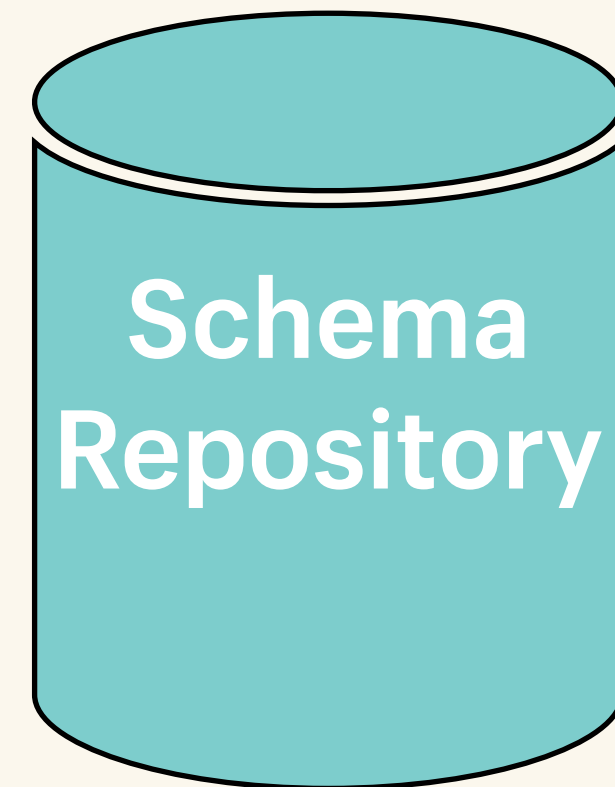
Data Subject	Controller	PII	Token
hooman@gmail.com	<i>allbirds</i>	hooman@gmail.com	token123
hooman@gmail.com		hooman@gmail.com	token456
hooman@gmail.com		 222-333-4444	Token789
bob@gmail.com		IP 76.44.55.66	Token011

+Retention

Analytical Events



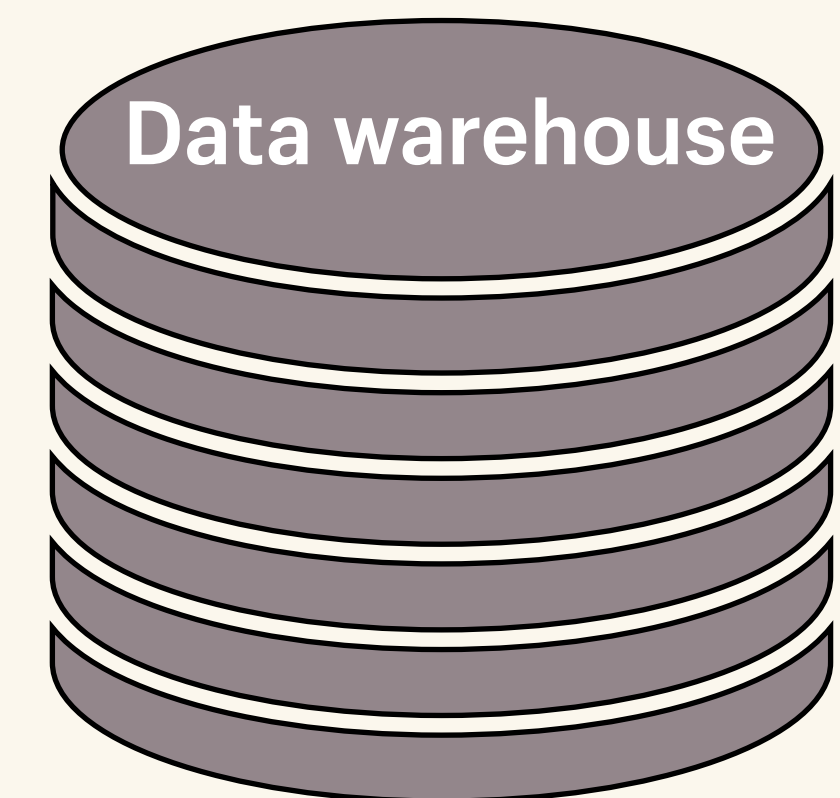
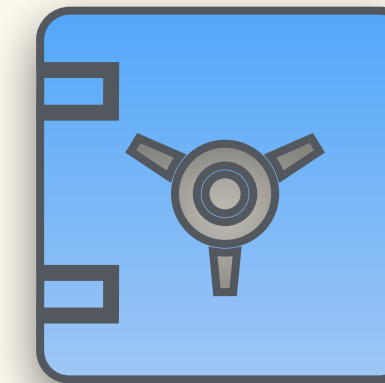
...



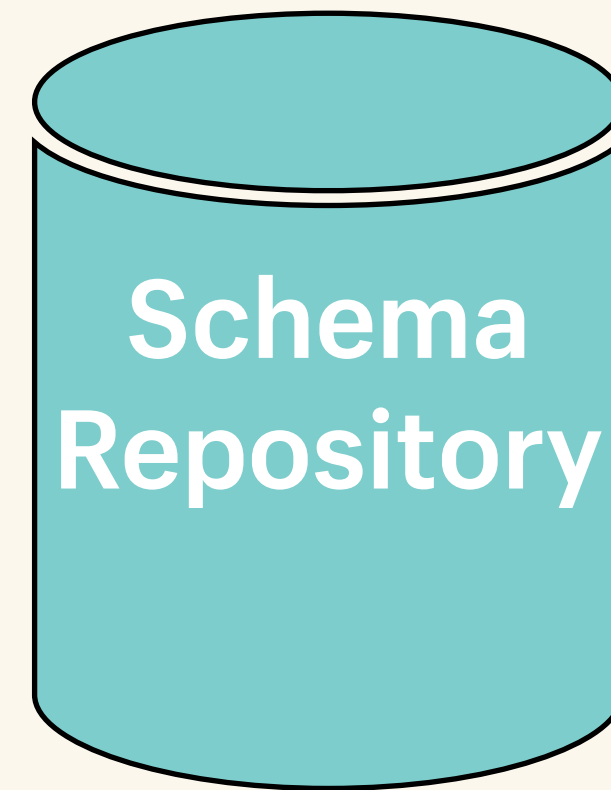
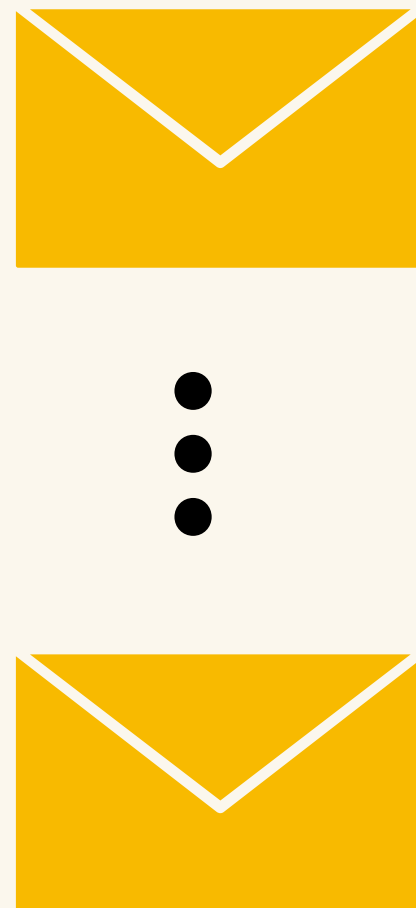
Scrubber



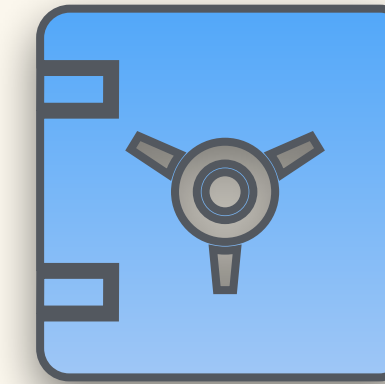
Tokenization
Vault



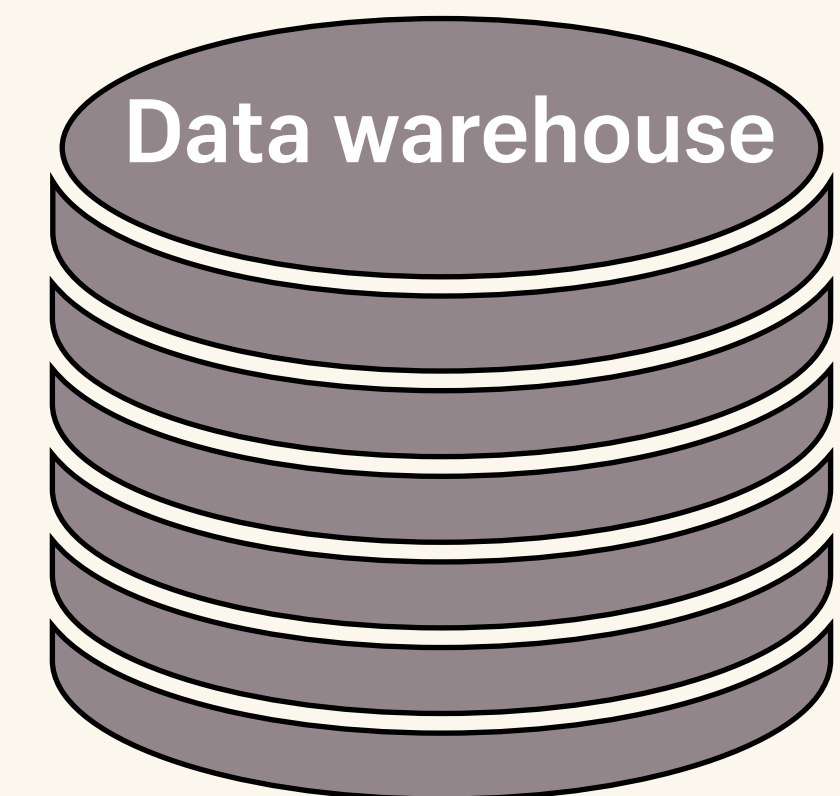
Analytical Events



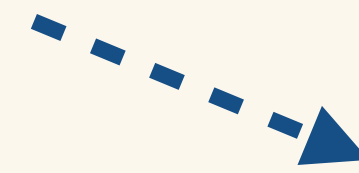
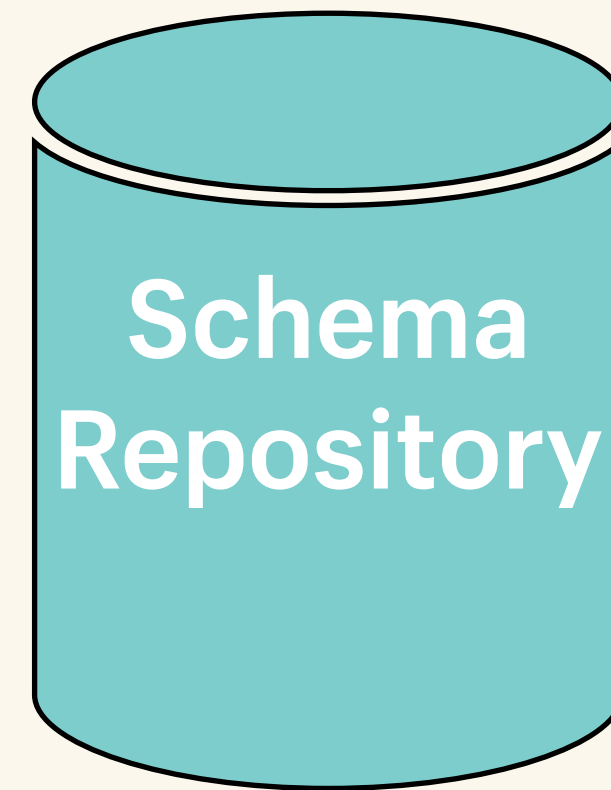
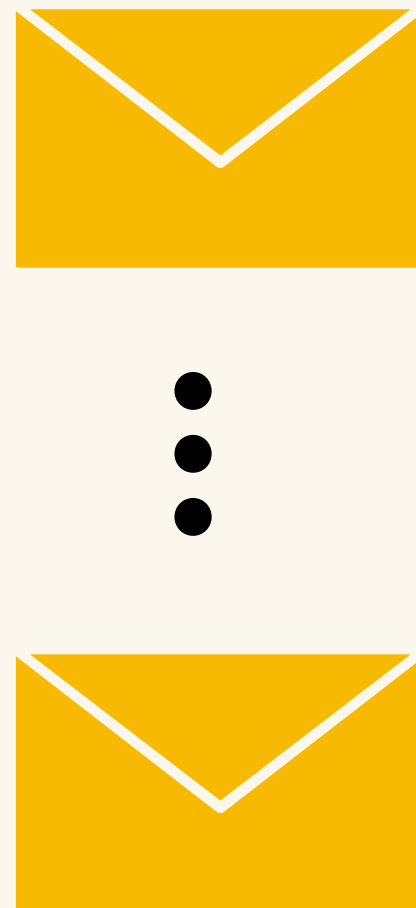
Tokenization
Vault



Scrubber



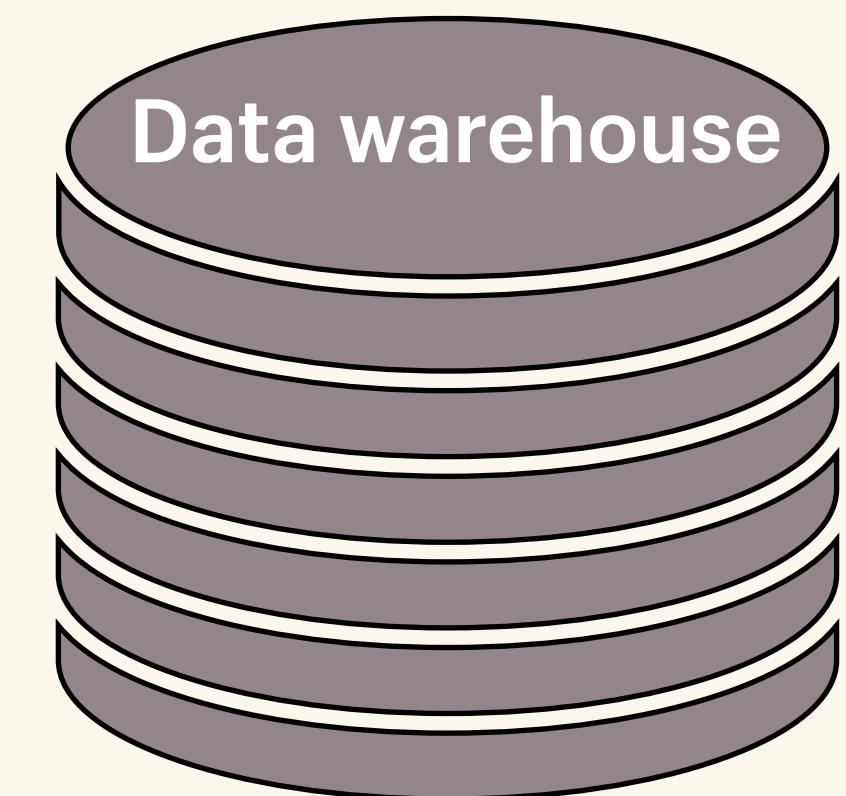
Analytical Events



Scrubber



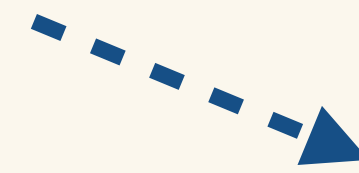
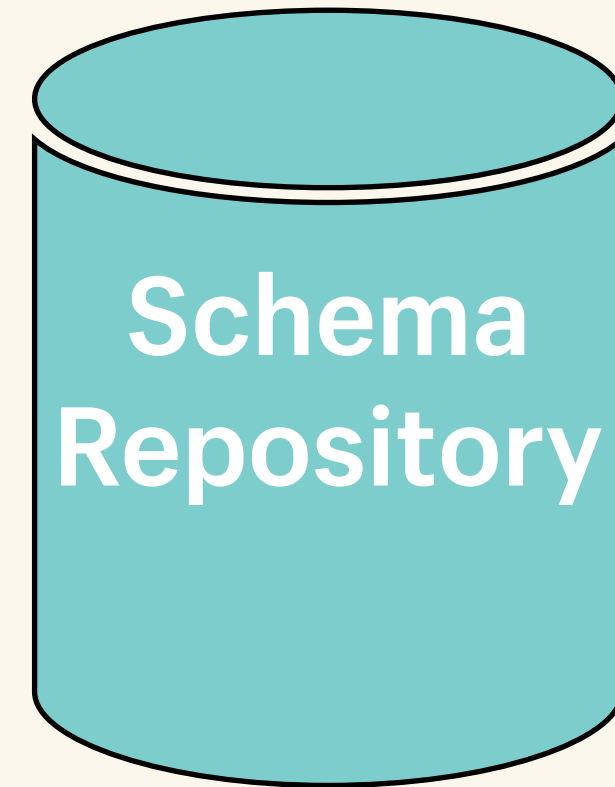
Tokenization
Vault



Analytical Events



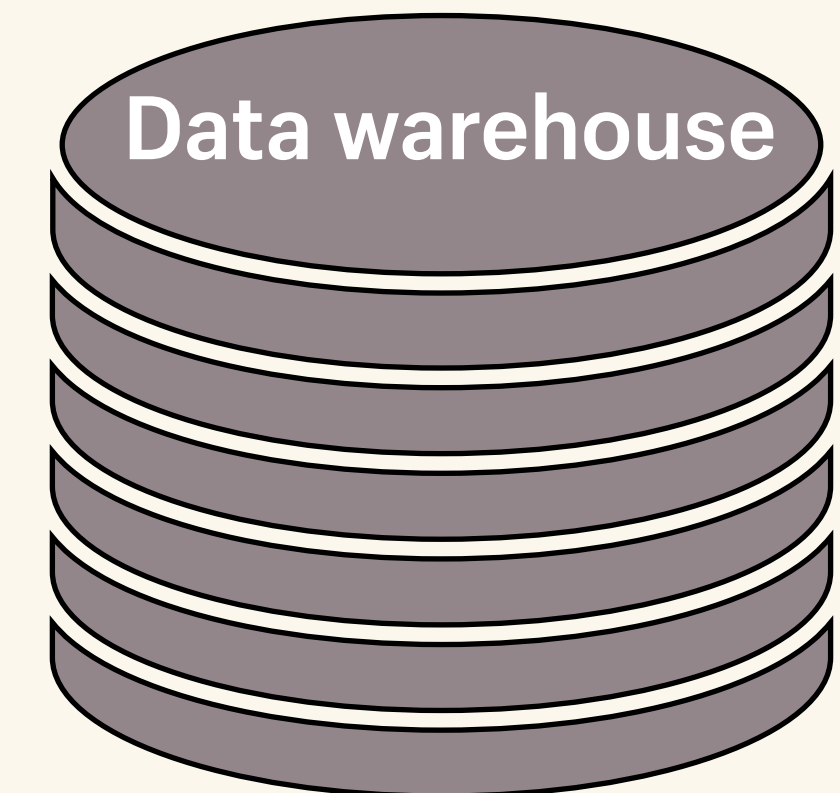
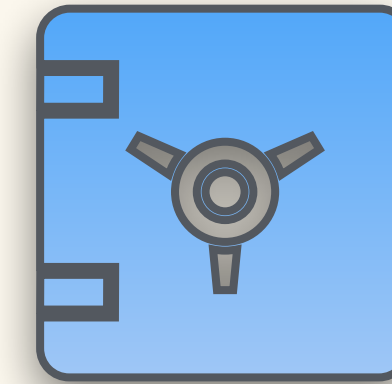
...



Scrubber



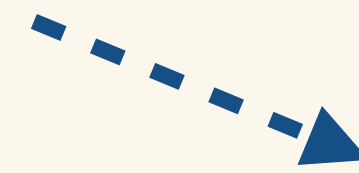
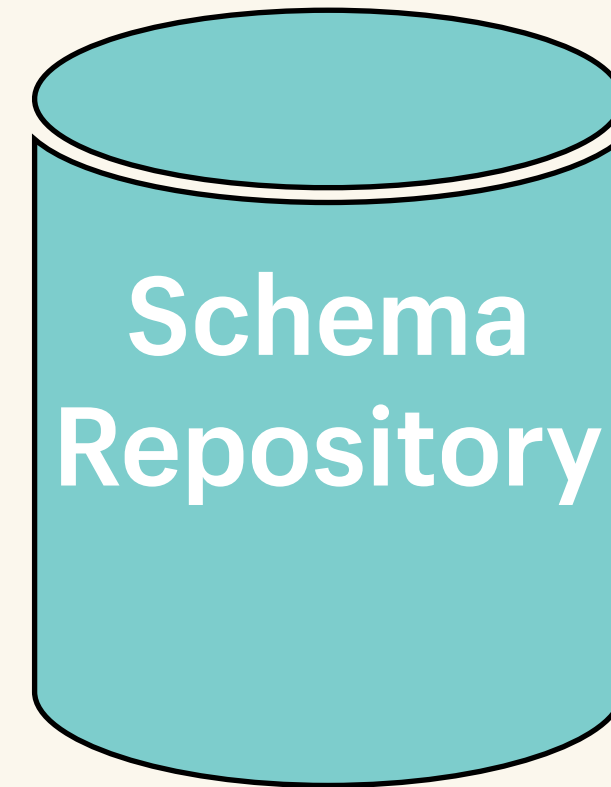
Tokenization
Vault



Analytical Events



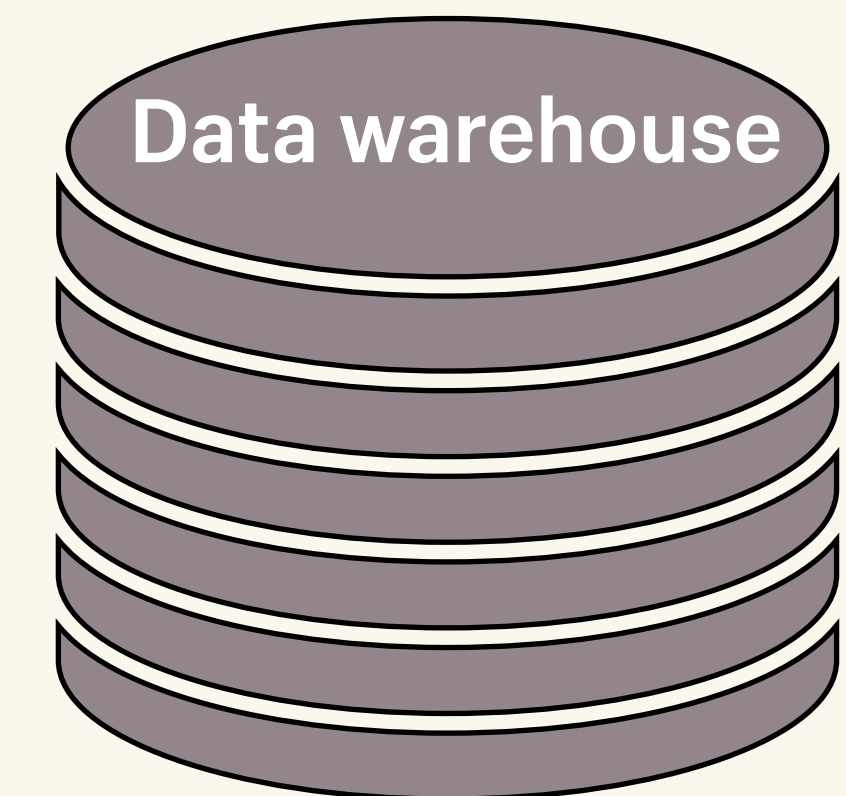
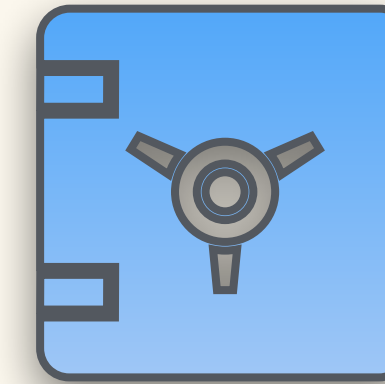
...



Scrubber



Tokenization
Vault



Are we good yet?

Challenges

- Adoption

Challenges

- Adoption



Challenges

- Adoption
- Addressing historic datasets



Adoption

- “*Trust me it will work*” is not enough!
- Make the wrong thing the hard thing
- Bring new value

Addressing historic datasets

- Organizational alignment
- Dedicated task force
- Tooling, documentation and support

2 Years Operation

0.5-1M

Event/Sec

10-100M

Deletions/Day

3Y

Dev/Rollout/Reprocess

500B

PII

100 (10P)

Historic datasets
reprocessed

94

Contributors

Summary

- Unstructured data is evil
- Tokenization/obfuscation
- Technical Challenges **≠** Organizational Challenges

Thanks

bshafiee@uwaterloo.ca
[@behroozshafiee](#)

