Data Masking Algorithms Strength

WHAT WE ARE GOING TO COVER

Prior Webinars:

- SSIS –first experience,
- Lifecycle,
- Testing with SSIS,
- ETL vs ELT,
- Data Masking Algorithms

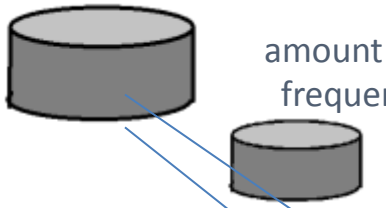
Substitution , Shuffling, Random , Date aging AKA Number and Date Variance, Encryption (yes, specific kind of encryption AKA tokenization), Nulling out or deletion, Value changes in increments or decrements, Custom

Today:

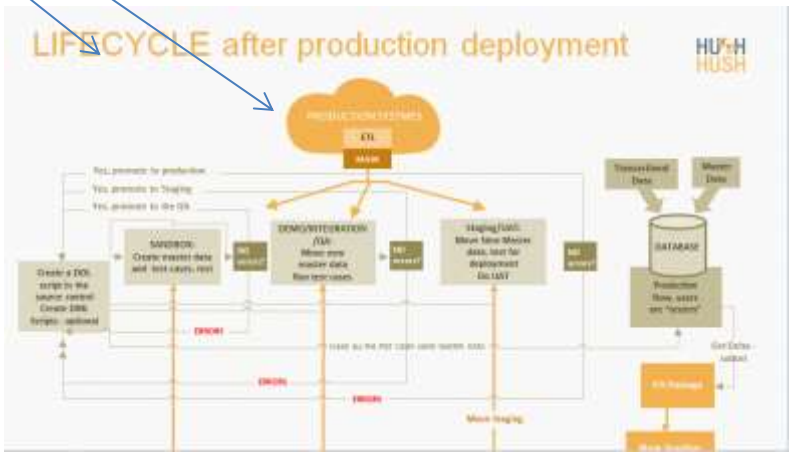
- Strength of different algorithms:
  - Shuffling
  - Substitution

PRIOR WEBINARS

testing techinques



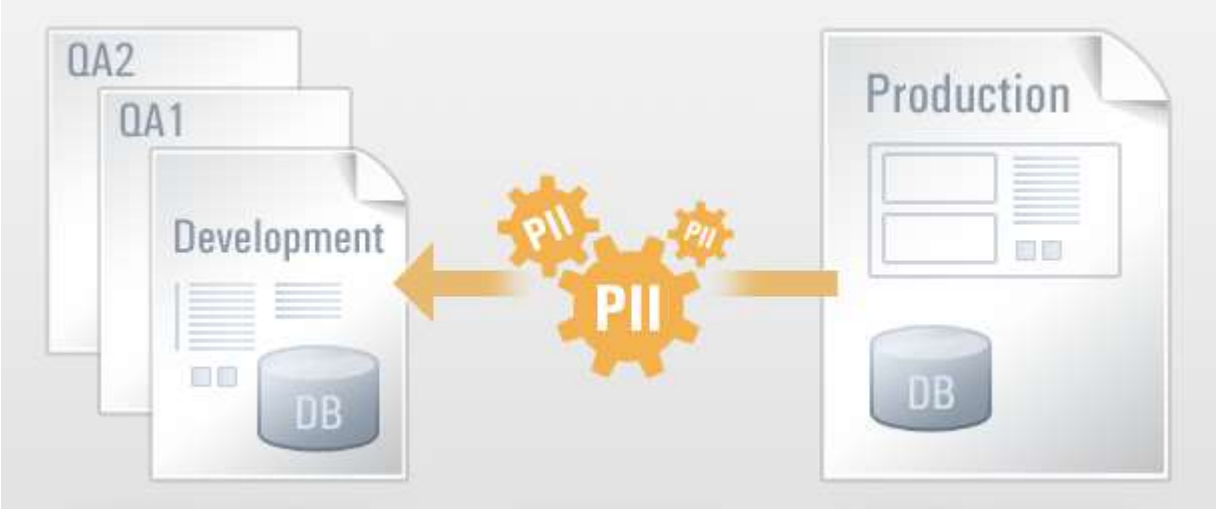
amount of data
frequency of uploads
SLAs , test coverage



data lifecycle

ETL vs ELT
With Delta and backup/load

DATA MASKING PUPROSE



Obfuscating sensitive data elements with “false” values within data store while preserving data look and feel and usability in applications.

HACKER-PROOFNESS

All animals are equal,
but some animals are more equal than others

"Value changes in increments or decrements" VS "FPE"

```
SELECT PopulationWithIncome as OriginalPopulation, PopulationWithIncome-201 as MaskedPopulation
FROM [dbo].[IncomeReportPerZip]
```

SHIFT 21

100 %

Results | Messages

	OriginalPopulation	MaskedPopulation
1	16769	16568
2	29049	28848
3	10372	10171
4	5079	4878
5	14649	14448
6	1263	1062
7	741	540
8	3609	3408
9	1370	1169
10	661	460

```
/****** Script for SelectTopNRows command from SSMS  ******/
SELECT TOP 10 [Phone].[Phone] as Phone, [MaskedPhone].[Phone] as FTP_Phone
  FROM [dbo].[Phone] INNER JOIN [dbo].[MaskedPhone]
  ON [Phone].[PhoneId] = [MaskedPhone].[PhoneId]
```

100 %

SHIFT

????

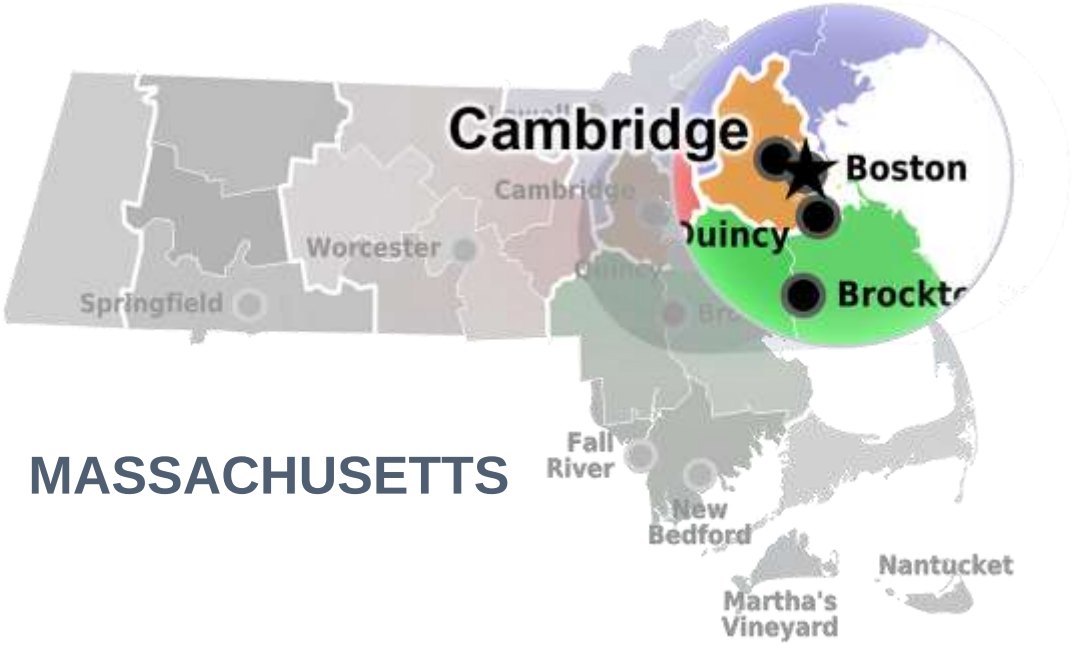
Results | Messages

	Phone	FTP_Phone
1	213-531-1853	606-667-6880
2	323.654.4567	567.266.9170
3	(712)666-9084	(994)386-1005

“SAFE HARBOR” - WHY HIPAA ASKS 18 ELEMENTS?

How difficult is it to find William Weld who lives in Cambrige, MA, city of 54,000 residents with 7 zip codes?

ZIP,
DOB,
GENDER,
PUBLIC DATA



WILLIAM WELD
MEDICAL
HISTORY

Only six people in Cambridge share the governor’s birth date, only three of them men, and of them, only he lives in his ZIP code. 87 percent of all Americans could be uniquely identified using only three bits of information: ZIP code, birthdate, and sex.

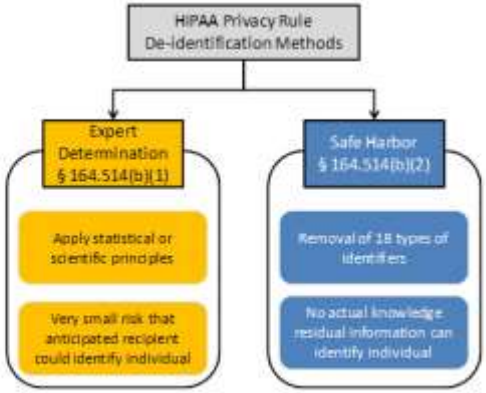
CHALLENGING REQUIREMENTS IN ALGORITMS IMPLEMENTATION

SECURITY

A NON-FUNCTIONAL REQUIREMENT → A FUNCTIONAL REQUIREMENT

- ❑ GLBA
- ❑ FERPA
- ❑ HIPAA

SAFETY REQUIRED IN IMPLEMENTATION:
“SAFE HARBOR”



. Two methods to achieve de-identification in accordance with the HIPAA Privacy Rule.

ALGORITHMS AND THEIR STRENGHTS

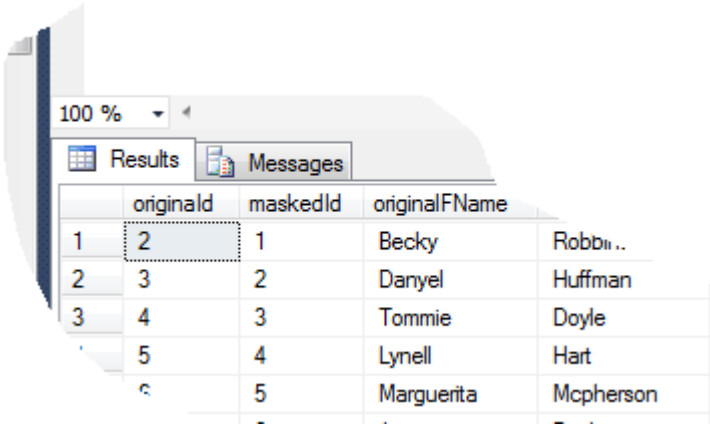
SHUFFLING

Definition: the data is shuffled within the column

Strength: great for maintaining aggregates.

Example: maintain the end of year figures for your financial information in the test data base

Different algorithms used for shuffling



	originalId	maskedId	originalFName	
1	2	1	Becky	Robb...
2	3	2	Danyel	Huffman
3	4	3	Tommie	Doyle
	5	4	Lynell	Hart
		5	Marguerita	Mcpherson

EXAMPLE OF SIMPLE SHUFFLE 1

Sample Production Database, #50

Last names and gender are not unique
Phone, and Phone# are unique

№	Last Name	Phone	License Plate	Gender
…				
45	Brown	3362548	33-88 a	m
46	Grey	4388753	54-21 a	f
47	Steinbeck	7355434	46-13 b	f
48	Smith	9987321	31-06 b	m
49	Orwell	7644321	54-12 a	f
50	Smith	8654443	32-43 b	m

Sample Masked Database with Shuffle,#50

Shuffle is done to Phone, License Plate and Gender

Shuffle: +3 +4 +2

№	Last Name	Phone	License Plate	Gender
…				
45	Brown	9987321	54-12a	f
46	Grey	7644321	32-43 в	m
47	Steinbeck	8654443	66-13 в	f
48	Smith	5434212	36-08 в	m
49	Orwell	6349071	12-23 a	f
50	Smith	3241112	32-12a	m

HOW DO WE HACK SHUFFLING?

The GOAL is to restore the table by known value.
Enter a row with unique value
Change TH to TTH in “Smith”

Sample Database, Edited, #51

We added new record with last name “Smitth” into production

№	Last name	phone	License plate	gender
…				
45	Brown	3362548	33-88 a	m
46	Grey	4388753	54-21 a	m
47	Steinbeck	7355434	46-13 в	f
48	Smith	9987321	31-06 в	m
49	Orwell	7644321	54-12a	f
50	Smith	8654443	32-43 в	m
51	Smitth	9090066	11-22 a	f

Sample Masked Database, #51
With Shuffle

№	Last name	phone	License plate	gender
…				
45	Brown	9987321	54-12a	f
46	Grey	7644321	32-43 в	m
47	Steinbeck	8654443	11-22 a	f
48	Smith	9090066	36-08 в	m
49	Orwell	6349071	12-23 a	f
50	Smith	3241112	32-12a	m
51	Smitth	4376542	25-09 в	m

HOW DO WE HACK SHUFFLING?

HUSH HUSH

Easy to establish shuffle value for unique fields

- ✓ phone (3)
- ✓ license plate (4)

Impossible to define Gender's shuffle

The probability for gender is 1/F, where F is amount of females

If the hacker knows the amount of females vs males, he will add the less populous gender.

Let's assume 20 females, then probability to recover gender field values is:

P=1/(# females+1)=1/21

1:1 SUBSTITUTION

Definition: Another authentic looking value can be substituted for the existing value.

Best for non-unique data

Substitution for unique SSN:
http://www.mssqltips.com/sqlservertip/3091/masking-personal-identifiable-sql-server-data/

Substitution for non-unique names:
http://www.brentozar.com/archive/2011/09/how-do-you-mask-data/

Results | Messages

	PositionNumber	OriginalText	MaskText
1	1	0	8
2	1	1	7
3	1	2	6
4	1	3	5
5	1	4	4
6	1	5	3
7	1	6	2
8	1	7	1
9	1	8	0
10	1	9	9
11	2	0	9
12	2	1	0
13	2	2	2
14	2	3	6
15	2	4	4
16	2	5	7
17	2	6	5
18	2	7	1
19	2	8	3
20	2	9	8

IMPLEMENTATION OF NAMES 1:1 SUBSTITUTION (as in a blog)

	First Name	First Name
1	Rubie	Angela
2	Esperanza	Annamae
3	Sunny	Ronnie
4	Nelida	Marco
5	Kacey	Georgeanna
6	Marisol	Etsuko
7	Hilaria	Oralia

"Ideally, the obscured data needs to have a similar distribution as the original data. If I have a People table with 1,000 records, all of which have a last name of Smith, then my obscured data should all also have the same last name. However, if my People table has 500 people named Smith and then 500 other people with unique last names, my obscured data needs to have 501 unique last names as well, one of which will have 500 records in the table."

In this method, we get distinct data set and create a substitution data set

Create 1:1 matching values table

HACKING NAMES 1:1 SUBSTITUTION

Sample Database of Diagnoses:

Data is evenly spread across States and represented in Proportion
1000 records

№	Surname	state	gender	medical diagnosis
1	BROWN	NY	m	
2	COOKE	CA	m	
3	GONZALEZ	NM	f	
4	HERNANDEZ	LA	f	
5	JOHNSON	CA	m	
6	KEY	CA	m	
7	MILLER	WA	f	
8	SMITH	NY	f	
9	WILLIAMS	KS	m	
	....			

We use 1:1 mapping table
To mask values

original surname	masked surname
BROWN	PARTIN
COOKE	NEMETH
GONZALEZ	ALMONTE
HERNANDEZ	PAN
JOHNSON	RICKARD
KEY	WENTWORTH
MILLER	SAMMONS
SMITH	SAYRE
WILLIAMS	SOUTHERLAND
....	

Resulting table with values:

№	masked surname	state	gender	medical diagnosis
1	PARTIN	NY	M	
2	NEMETH	CA	M	
3	ALMONTE	NM	F	
4	PAN	LA	F	
5	RICKARD	CA	M	
6	WENTWORTH	CA	M	
7	SAMMONS	WA	F	
8	SAYRE	NY	F	
9	SOUTHERLAND	KS	M	
	….			

Site with name frequencies: http://names.mongabay.com/

Masked data has the same frequencies as original:

original surname	Occurrences per 100,000 people	approximate occurrences per 10,000 people	masked surname
BROWN	511.62	51	PARTIN
COOKE	11.81	1	NEMETH
GONZALEZ	221.57	22	ALMONTE
HERNANDEZ	261.85	26	PAN
JOHNSON	688.44	69	RICKARD
KEY	11.82	1	WENTWORTH
MILLER	418.07	42	SAMMONS
SMITH	880.85	88	SAYRE
WILLIAMS	568.66	57	SOUTHERLAND
….			

SUBSTITUTION 1:1 FREQUENCES

the frequency of a substitution for the last names JOHNSON will coincide with the frequency of last name JOHNSON

the second most frequent name

There will be 69 last names in the database with 10000 records

original surname	frequency	masked surname
SMITH	88	SAYRE
JOHNSON	69	RICKARD
WILLIAMS	57	SOUTHERLAND
BROWN	51	PARTIN
MILLER	42	SAMMONS
HERNANDEZ	26	PAN
GONZALEZ	22	ALMONTE
COOKE	1	NEMETH
KEY	1	WENTWORTH

№	masked surname	state	gender	medical diagnosis
1	RICKARD	NY	m	
2	RICKARD	CA	m	
3	RICKARD	NM	f	
4	RICKARD	LA	f	
5	RICKARD	CA	m	
6	RICKARD	CA	m	
7	RICKARD	WA	f	
69	RICKARD	KS	m	

SUBSTITUTION 1:1 FREQUENCES

CHANCE TO CORRECTLY ESTABLISH THE DIAGNOSIS BASED ON NAMES ALONE: 1/69

ADDING STATE :

2011 Census

California population: 38 million people'

Total United States population: 310 million people

Probability of being a Californian: 38/311 = 0.122

among the 69 RICKARD will be approximately 12% of Californians, about 8 people

Probability of "RICKARD/CALIFORNIAN" 1/8

Considering Gender: 1.4

NOT THE BEST SECURITY

SUBSTITUTION DISTURBING STATISTICS

Probabilities increase when we don’t do 1:1 substitution and increase the disturbance in to the statistics. There is a lot of work being done in this area in different research labs such major universities , Microsoft and other companies. We translate this research it into the product- so that you do not have to.

CONCLUSION

CREDITS

Lyudmila Kirichenko, the Doctor of Mathematics and Statistics - expertise and examples
Yuriy Lobzakov, HushHush algorithms architect - expertise
Aviva Sterns – voice
Andrew Coyne – producer and editor
Virginia Mushkatblat – script writer and director

HUSH HUSH

Hush Hush

Virginia Mushkatblat

http://mask-me.net

Mobile: 213.631.1854

Office: 855.YOU.HUSH

info@mask-me.net