HushHush
848 N. Rainbow Blvd 2890, Las Vegas, NV, 89107
1-855-YOU-HUSH Info@mask-me.net

White Paper
HushHush Data Masking

Anonymization compliance to the laws of the European Union and North America

EU

The laws of the European Union do specify that data should be anonymized and/or pseudonymised, in Convention 108, Article 5 (e) Convention 108, Explanatory report, Article 42. In particular, data must be kept "in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the data were collected or for which they are further processed." For that, data has to be anonymized straight after use and archived. The definition of "necessary" also leads to pseudoanonymization although not directly mentioned in such aspect. However, in continuous development of the applications, there is no direct need for the developers, for example, to see sensitive data. Also, sometimes the further need arises to use archived data and for keeping meaning of the complete context, not just deleting data out of the context.

USA and Canada

USA and Canada have a variety of laws serving financial, healthcare, educational, government industries in terms of privacy regulations. Oftentimes, data masking is the answer to complex requirements to limit access to PII. HIPAA serves as a particular example, with two methods listed acceptable for data de-identification, Safe Harbor and Expert Determination. Safe Harbor lists particular elements to hide.

As per above laws, HushHush components established in SSIS satisfy both requirements. The tool allows both data deletion and data de-identification in structured and semi structured storage. The tool also can allow rudimentary scrambling of text data, and if employed with a parser, de-identification of the sensitive elements of text.

As the complexity of the recognizing semantic closeness between entities is NP-complete, **any algorithmically informed vendor** will **answer** of the question of " whether they provide compliance out-of-the-box" as either "sufficient" or "partial" – to maintain scientific objectivity. However, HushHush does provide "out- of-the-box compliance" with "out-of-the-box components" on par or exceeding capabilities of the rest of the vendors. These components envelop algorithms that contain not only data but also account for the statistical variance of non-unique data elements.

Anonymization performance.

Statistical properties of the tool.

The guarantees to the compliance can be defined in terms of the variability of the domain of identifying a persona from her/his semantic properties mapped in the application domain.

In simpler words, when developers create some data architecture, they capture certain traits and states of the individual and her/his events/behavior/actions into words. These words become metadata and data in the application's different formats of storage, and create the data domain.

It is from the **data** in such domain that **the person can be identified with**. The core of the domain based on the protection from the known threats was identified by the Latanya Sweeny for the medical applications in the 18 rules of HIPPA (http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/guidance.html#_ednref16) . These rules pertain to the domain of the medical community but could be sufficiently generalized. These specific rules were taken as a guidance when creating a tool. They provide for Safe Harbor requirements and more. Specific components account for statistical properties of data both with their data collections and algorithms. Such are geographical elements, mainly zip codes, and dates. Every single component accounts for whether the element should be unique or has statistics of known occurrences to account for within data collection, and if used properly will provide the sufficient guarantee. In the world of applications, there are also elements of the data domain not mentioned in the core – and these are accounted with the common components that allow to create necessary statistical properties and use generalized algorithms. The tool in particularly easily accommodates to specifics of any additional entities, so under the European Law it is flexible enough to adopt to the domain quickly.

Such patterns as generalization (e.g. zip codes 10001-10009 are generalized with 1000*) are easily accomplished with regular SSIS dynamic expression engine techniques. There is an understating in the scientific community and industry (as the very Safe Harbor article indicates) that the **guarantees** are indeed **in the recognition of all the necessary entities and necessary algorithms,** and as mentioned in the paragraph 1, it is still an NP-complete task requiring human's evaluation, thus **HushHush specific components are compliant to the core of the domain** identified by Safe Harbor, to the extent at all possible and has **generic algorithms** satisfying Expert Determination method of HIPAA and used for other types of compliance such as GLBA and PCI/DSS.

Tool Philosophy

HushHush is not only an algorithm and services provider but by necessity also the data provider. There are a lot of the algorithms that are based on the statistical sets available in the public domain and these sets require maintenance.

Tool Differentiators

- HushHush allows for easy integration into the enterprises that have established SQL base
- HushHush allows for performance tuning not available with "boxed" tools and it is a significant consideration for organizations with large data volumes
- HushHush easily extends: in the case of the particularly European Union requirements, it is possible to add components satisfying the specifics of the law's domain.
- HushHush partners with data quality tools Components Company and as such complements the data lifecycle of the company to the completion, with both its master data and transactional data.

Anonymization properties of HushHush.

Comparison to scientific anonymization models.

The tool can provide at least k-4 anonymity per table and l-diversity $l \geq 3$.

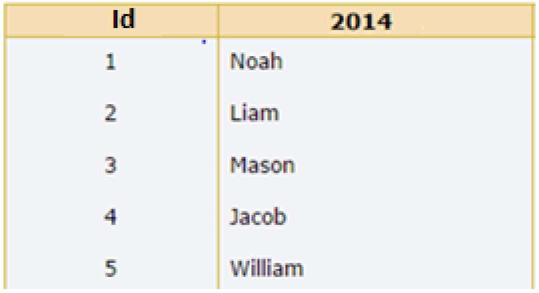
See example below.

Let's consider below subset of medical records as our sample.

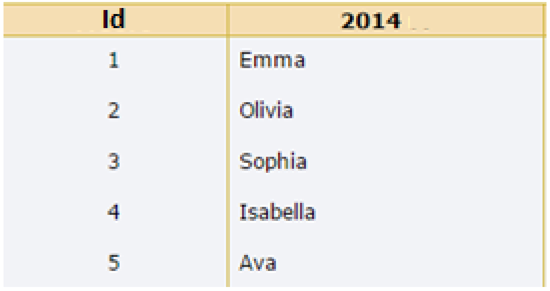
Name	Gender	Zipcode	Disease
Andre	Male	53715	Heart Disease
Beth	Female	53706	Hepatitis
Carol	Female	53703	Bronchitis
Dan	Male	53703	Broken Arm
Ellen	Female	53716	Flu
Eric	Male	53705	Hang Nail
Beth	Female	53707	Acne
Andre	Male	53712	Heart Disease
Eric	Male	53704	Heart Disease
Andre	Male	53716	Flu
Beth	Female	53703	Heart Disease
Ellen	Female	53712	Cancer
Holly	Female	53714	Heart Disease

Ben	Male	53712	Cancer
Ellen	Female	53715	Cancer
Dan	Male	53704	Bronchitis

We will de-identify this sample with the help of the tool's algorithm (patent pending). This de-identification method uses most common US names for males and females according to Census.

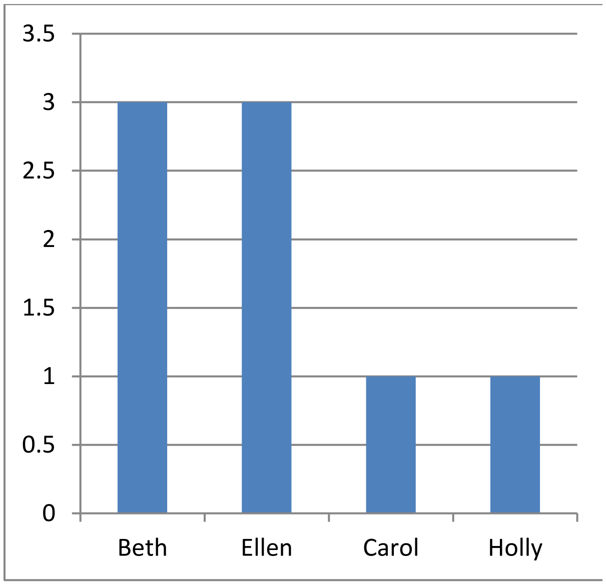


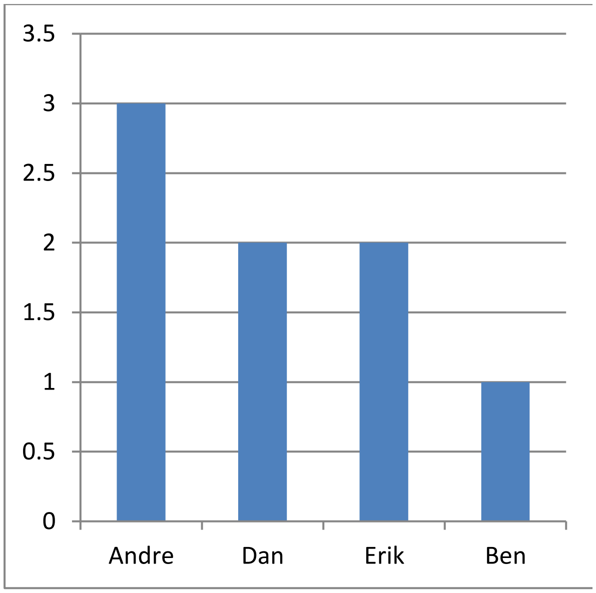
Id	2014
1	Noah
2	Liam
3	Mason
4	Jacob
5	William



Id	2014
1	Emma
2	Olivia
3	Sophia
4	Isabella
5	Ava

The frequencies of the population occurrence are presented at the diagrams below.



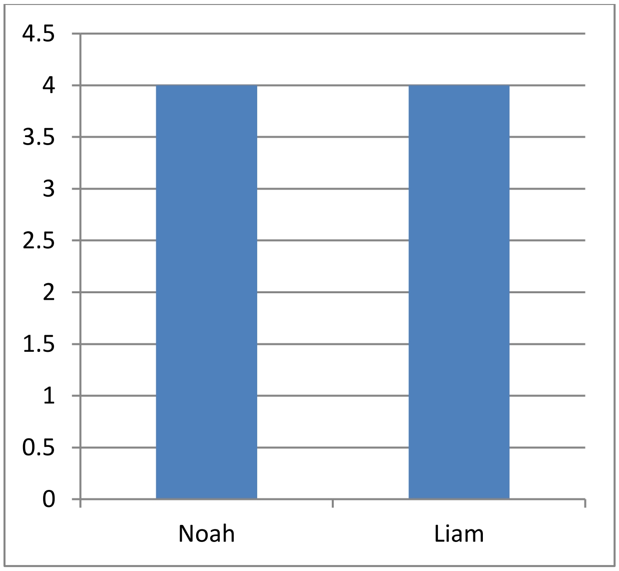


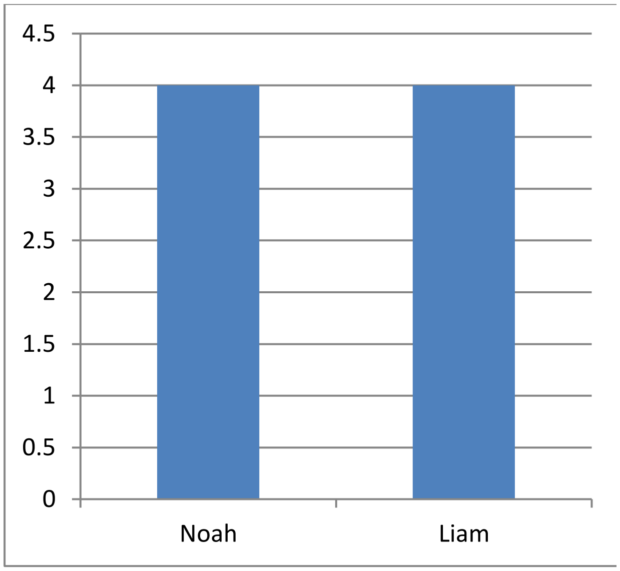
		frequency
Beth	3	0.375
Ellen	3	0.375
Carol	1	0.125
Holly	1	0.125

		frequency
Andre	3	0.375
Dan	2	0.25
Erik	2	0.25
Ben	1	0.125

Components' algorithms supposes the transformation of the source occurrence distribution pattern into the destination's uniform distribution pattern. The transform is illustrated below in the following tables and diagrams:

Name	Gender	Zipcode	Disease
Emma	f	10007	Bronchitis
Emma	f	10007	Heart Disease
Emma	f	10007	Acne
Emma	f	10007	Hepatitis
Liam	m	10007	Broken Arm
Liam	m	10007	Hang Nail
Liam	m	10007	Bronchitis
Liam	m	10007	Heart Disease
Noah	m	10008	Heart Disease
Noah	m	10008	Cancer
Noah	m	10008	Heart Disease
Noah	m	10008	Flu
Olivia	f	10008	Cancer
Olivia	f	10008	Flu
Olivia	f	10008	Cancer
Olivia	f	10008	Heart Disease





Bath +Carol	0.375+0.125	Emma	0.5
Ellen+Holly	0.375+0.125	Olivia	0.5

Andre+Ben	0.375+0.125	Noah	0.5
Eric+Dan	0.25+0.25	Liam	0.5

We are not masking the gender field in this particular scenario as the majority of names give out the gender of the persona.

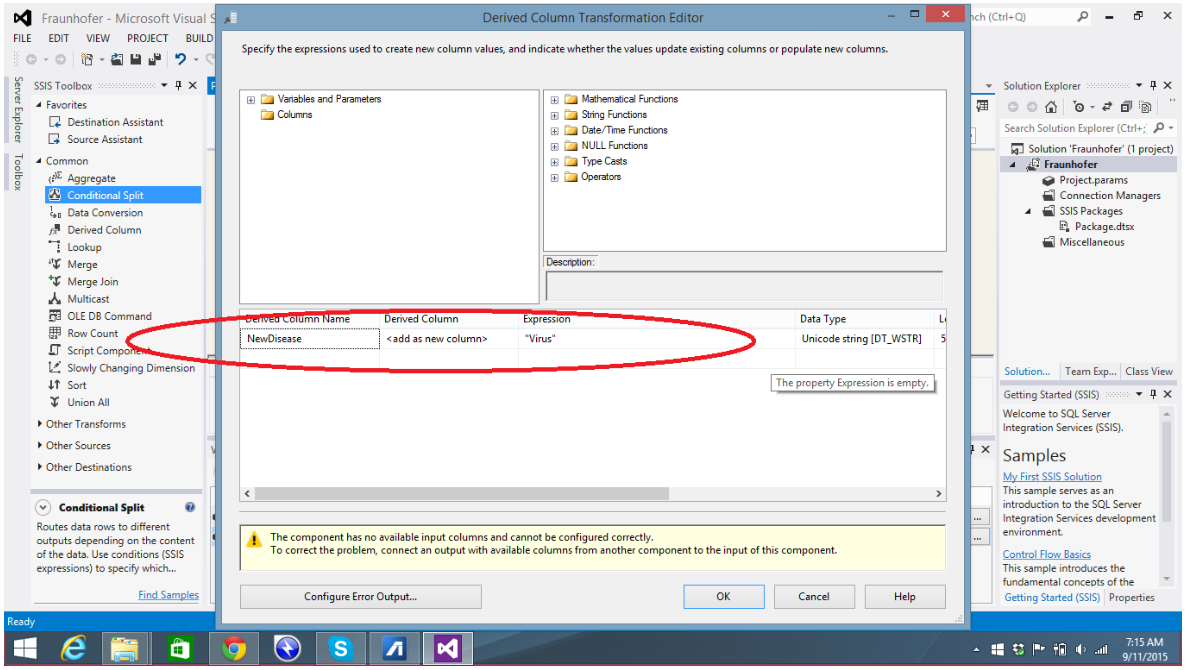
In order to mask the zipcode field, let's break data into two groups, as there is a pretty big variability of zipcodes, of which four zipcode occur only once. Zipcodes exceeding 53710 we will substitute with 10008 and those that are less or equal 53710 let’s substitute with 10007 (New York’s zip code).

Resulting table is 4 –k anonymous table as every masked value repeats no less than four times. Along with k-anonymity, this table satisfies requirements of l-Diversity as inside each group of records, $l \geq 3$.

This algorithm is effective for de-identification of sufficiently populated sets of records, as small sets are proven to be difficult to de-identify in principle.

Appendix One

Increasing L-diversity with Transformation editor of Derived component of SSIS



Derived Column Transformation Editor

Specify the expressions used to create new column values, and indicate whether the values update existing columns or populate new columns.

Derived Column Name	Derived Column	Expression	Data Type	L
NewDisease	<add as new column>	"Virus"	Unicode string [DT_WSTR]	5

The property Expression is empty.

The component has no available input columns and cannot be configured correctly.
To correct the problem, connect an output with available columns from another component to the input of this component.

Conditional Split

Routes data rows to different outputs depending on the content of the data. Use conditions (SSIS expressions) to specify which...

Find Samples

Welcome to SQL Server Integration Services (SSIS).

Samples

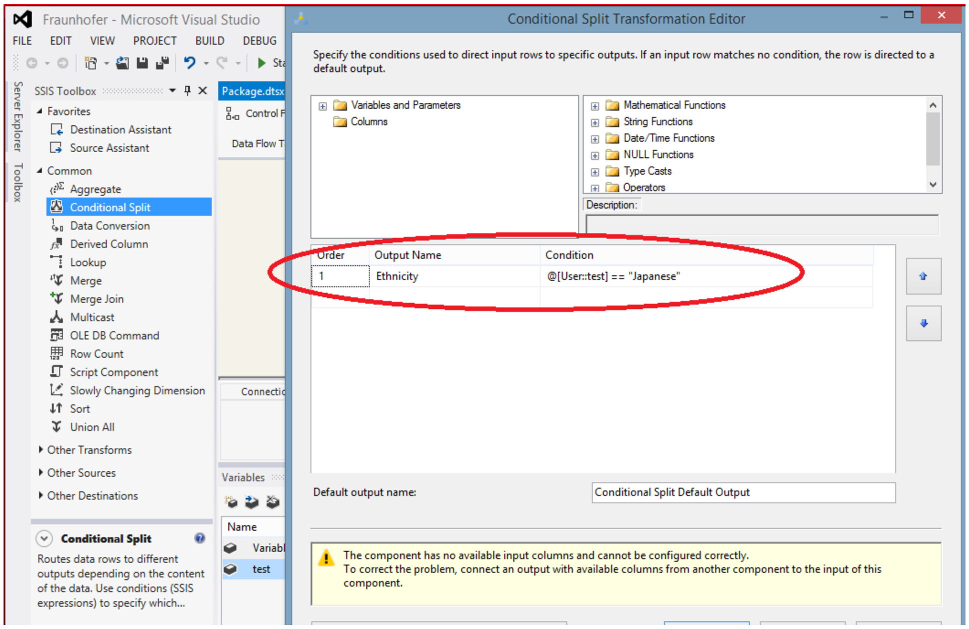
My First SSIS Solution

This sample serves as an introduction to the SQL Server Integration Services development environment.

Control Flow Basics

This sample introduces the fundamental concepts of the

Assigning values based on other attributes with conditional Split to increase k-anonymity



Conditional Split

Routes data rows to different outputs depending on the content of the data. Use conditions (SSIS expressions) to specify which...

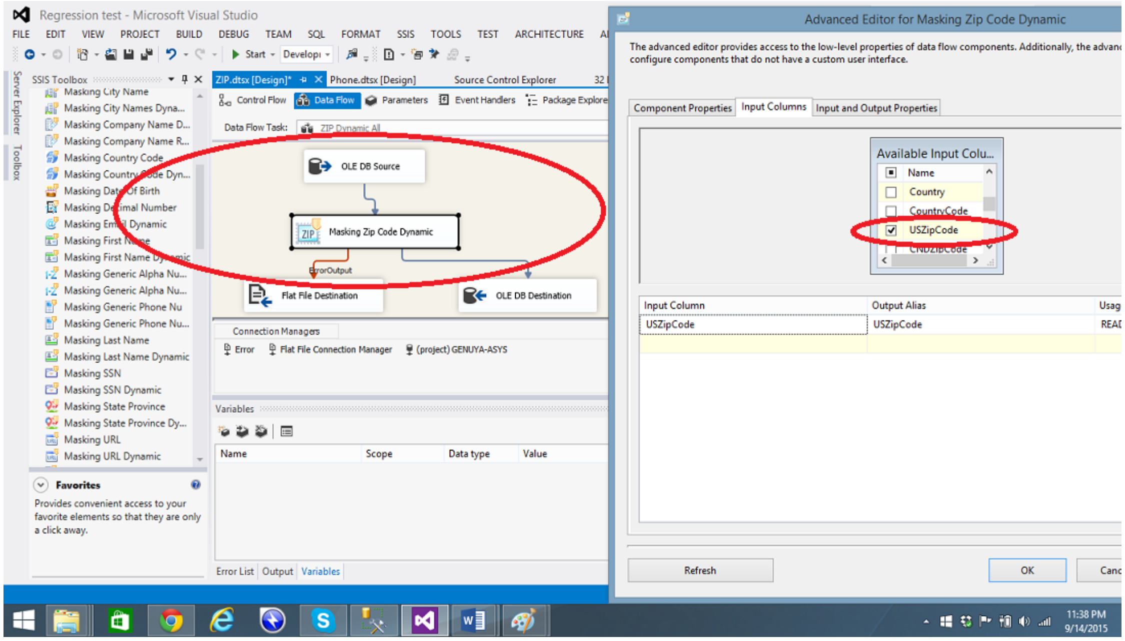
Specify the conditions used to direct input rows to specific outputs. If an input row matches no condition, the row is directed to a default output.

Order	Output Name	Condition
1	Ethnicity	@[User::test] == "Japanese"

Default output name: Conditional Split Default Output

The component has no available input columns and cannot be configured correctly.
To correct the problem, connect an output with available columns from another component to the input of this component.

Using out of the box Zip code algorithm and data set that automatically maps codes of the sparsely populated areas:



Provides convenient access to your favorite elements so that they are only a click away.

The advanced editor provides access to the low-level properties of data flow components. Additionally, the advanc
configure components that do not have a custom user interface.

■ Name
□ Country
□ CountryCode
☑ USZipCode
□ CNDZipCode

Input Column	Output Alias	Usag
USZipCode	USZipCode	READ