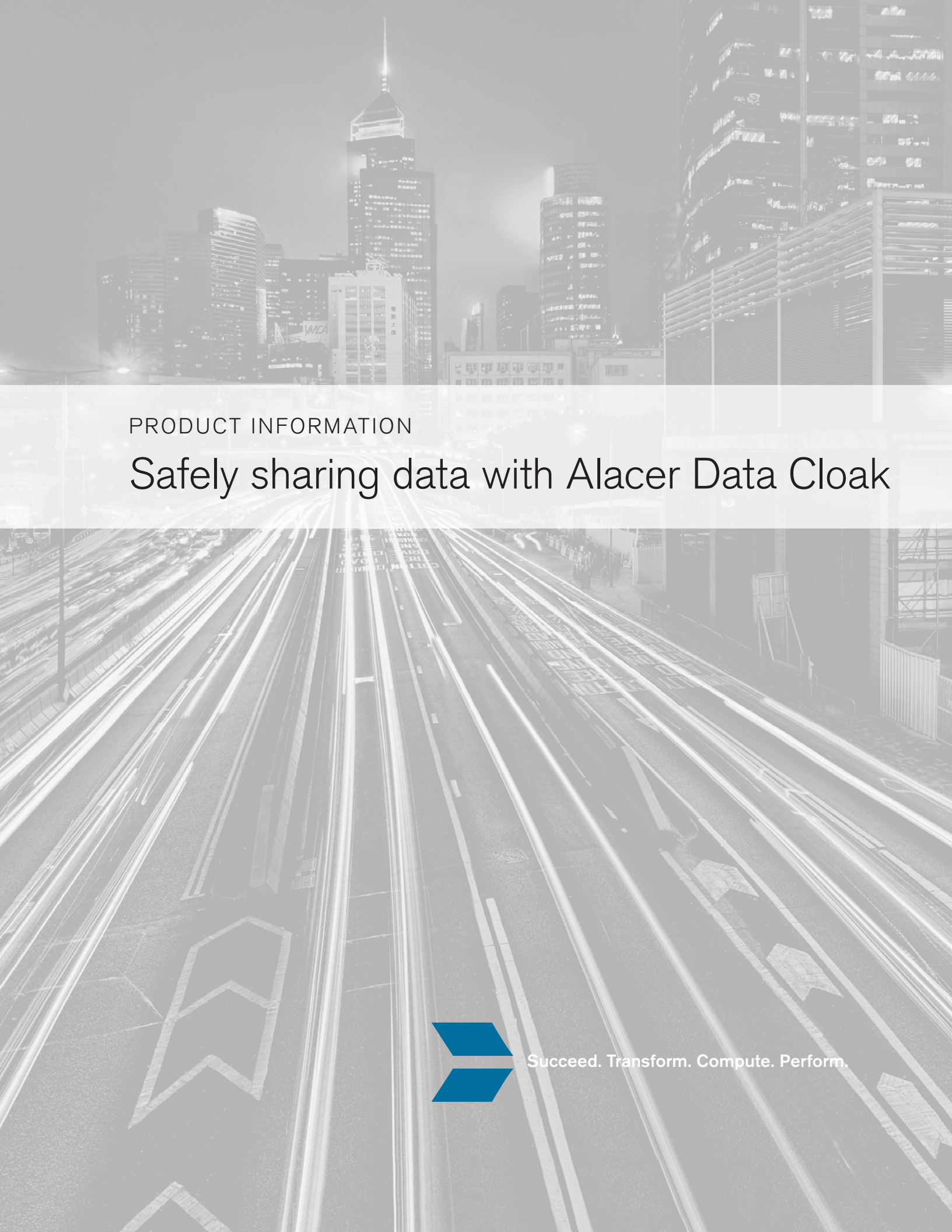


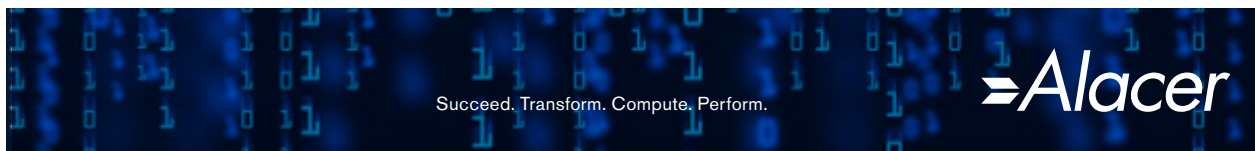


PRODUCT INFORMATION

Safely sharing data with Alacer Data Cloak



Succeed. Transform. Compute. Perform.



Safely sharing data with Alacer Data Cloak

PATENT-PENDING

BY AARON HOFFER, ALACER TECHNOLOGY SOLUTIONS

The data security paradox

Data gains value when it's available to the right audience. The greater the audience, the greater the potential for unlocking value in the data. However, it is often the most sensitive data that is also the most valuable. Customer records, transaction histories, and medical information are a rich source for data mining. These datasets are also closely guarded, and their confidentiality might be legally protected or their access restricted by corporate policy. The reasons are clear: a data break can be disastrous for a company and its customers. This is the data security paradox: the more valuable the data, the harder it is to obtain the data.

WHY IT MATTERS

First and foremost, the data security paradox affects the owners of the data. A bank department might want to analyze transaction data in a public cloud, like Azure or AWS, but are restricted from doing so. Or the bank department might want an outside analyst to look for money laundering activities, but are afraid the data might be leaked to the Internet.

Second, the data have value to an audience outside the data owner's organization and its trusted partners. Prevailing wisdom is that data can not be safely shared with the public. "Bad actors" could use the data maliciously or inappropriately. However, this audience also contains researchers, non-profits, NGOs, think tanks, and altruistically minded individuals can use the data to discover new insights, advance everyone's understanding of a subject, and create new value. A McKinsey Global Institute reports that "open data can help unlock \$3 trillion to \$5 trillion in economic value annually..." ^[1]

^[1] "Open data: Unlocking innovation and performance with liquid information," McKinsey Global Institute, October 2013. James Manyika, et. al. http://www.mckinsey.com/insights/business_technology/open_data_unlocking_innovation_and_performance_with_liquid_information

We set out to lift
the burden of
sanitizing data...

SAFELY SHARING DATA

Organizations share data, even sensitive data. Smart organizations lower their risk by anonymizing or obfuscating sensitive data, especially personally identifiable information (PII). PII includes identifying information like name, phone number, or date and place of birth. PII extends to anything that can identify who the person is, including medical health history or IP address. Getting data in a state where it is ready to share means removing PII and anything that might be damaging in the context of the data. This is sometimes called “sanitizing” the data.

A common approach to sanitizing is to remove all sensitive attributes of the data before sharing it. This approach has two potential problems. First, the sensitive fields are often the primary keys of the data. Once these keys are gone, the ability to connect sanitized records with the original data is lost. This can be important. For example, if a third party sanitized bank transactions to find money laundering, the data owners will ultimately want to know which customers are implicated. If all PII is simply excised, this could be difficult or impossible to do.

The second problem is that sensitive data can be intrinsically valuable. Sometimes this loss is unavoidable. For example, removing a customer's name from anti-money laundering data before sharing it would preclude an outside analyst researching that customer's legal filings. However, sometimes the value of the information can be retained even if the data is sanitized. To continue with the banking example, the transfer of money between customers can reveal important patterns, even if the names of the customers are not known.

...we created a browser-based application to transform...sensitive data into a public (shareable) file and a private file.

LIFTING THE BURDEN OF SAFELY SHARING DATA

The average analyst does not have the tools or experience to sanitize data. We have seen clients mitigate the problem by either never sharing data or sharing data after signing NDAs. Sometimes, the data are shared with nothing more than the hope that they are not compromised.

We set out to lift the burden of sanitizing data so they can be safely shared. We wanted to create an application that data analysts could use to remove sensitive information so that data could be shared with someone in another group at the same organization, an outside business partner, or even industry analysts and researchers. Important requirements for the solution included:

Spreadsheets: Most of the data our clients want to share is in Excel spreadsheets. Even if the data originates in a RDBMS, the data is typically exported as a flat file before being shared.

Accessible: Many data analysts work in secure environments where they cannot install applications on their computers. None of our potential users wanted to install an application and many would not trust an executable file even if they could install applications.

Simple: Complexity is a common source of security weaknesses. Simple systems are easy to analyze and their behavior is more predictable and less likely to contain bugs.

Isolated: The solution had to run entirely on the user's computer with no need to access the network. Security becomes much more difficult once data is transmitted over a network.

Irreversible: Data must be sanitized in such a way that it was impossible to recover sensitive information after it was anonymized or obfuscated.

Retain relationships: Relationship mining is an important part of many analyses. The solution had to preserve relationships in the original data.

Two-way: In our use cases, sanitizing data to make it ready to be shared is only half the equation. Once the data is shared, another party typically analyzes it and enhances it by adding additional data attributes. For example, an outside analyst might work with shared customer data to create a risk score for each customer. The analyst then provides that information back to the data owner. It should be easy for the data owner to join the enhanced data with the original, sensitive data.

The Alacer Data Cloak solution

To meet our requirements, we created a browser-based application that transforms a file with sensitive data into a public file and a private file. The shareable public file contains sanitized information. It is safe to share with outside parties. The private file contains the sensitive information and stays with the data owner. The private file can be used to re-integrate (uncloak) analysis provided by an outside party.

Browser-based application: The Data Cloak runs entirely in a browser. This makes the application very accessible; there is no need to install or start an executable. Browsers are installed on almost every workstation. Browsers also allow the application to run transparently on different operating systems (e.g. Windows, Mac, Linux).

Comma-separated value files: The Data Cloak operates on comma-separated value (CSV) files. All spreadsheet applications can save their output as CSV files. Spreadsheet programs can also import CSV files easily. CSV files are not specific to an operating system, browser or other technology.

User interaction: The idea behind Data Cloak is simple. The Data Cloak scans an input file and presents the user with a simple list of choices for each attribute in the data. The user easily selects one of three actions for each attribute:

- Declare the attribute is safe to **remain** in the shared data.
- Declare the attribute as sensitive and should be **cloaked**.
- Declare the attribute as sensitive and **remove** from the shared data.

Local processing: No data, not even sanitized data, are transmitted to another computer. The output is generated and saved to the user's local storage files. The sanitized public file can then be safely emailed to the intended recipient.

Random generation: Sanitized values are randomly generated. The sensitive data is not used as part of the sanitizing process. There is no way to divine the original data from their sanitized replacement values.

Global substitution: Sensitive values are always replaced by the same cloaked value to preserve relationships between entities.

Uncloak application: A second browser-based application, Alacer Data Uncloak, recombines the public and private files to uncloak or restore the data.

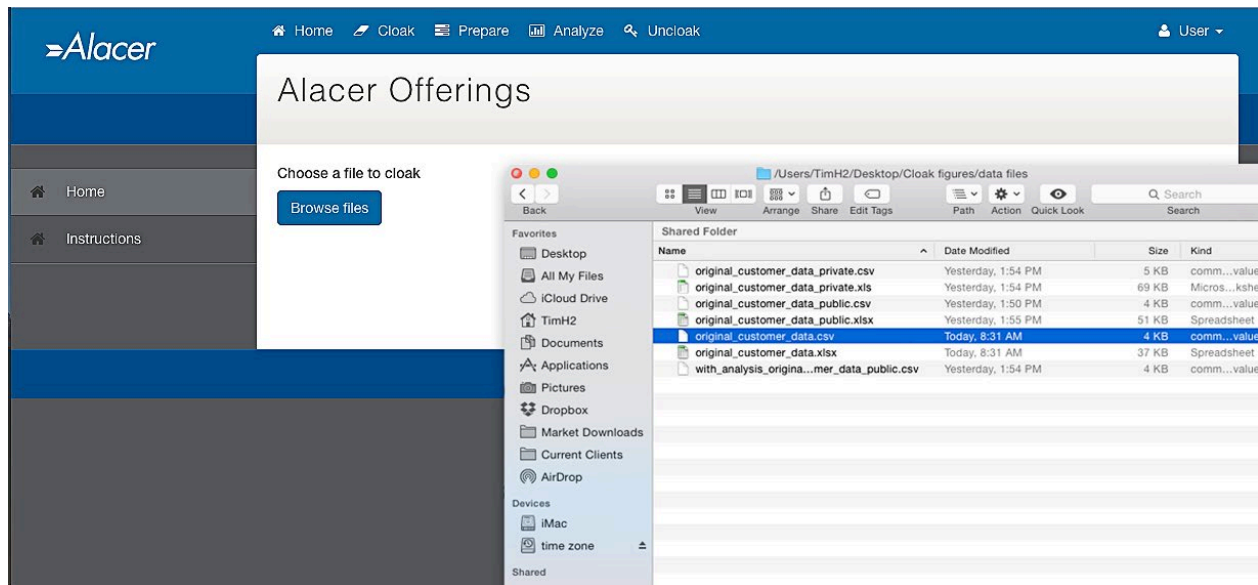
DATA CLOAK INPUT

The input for the data cloak is a delimited CSV text file, typically created using Microsoft Excel's "Save As" command. The table below shows the first few rows of the spreadsheet used in this example. It contains information about a group of customers, representative of anti-money laundering data.

CustID	CustomerName	SSN	Address	City	County	State	Zip	Country	Customer Type	Income Category	NAIC	Number Of Accounts
20130100001	MERRILL MEISENBURG	688297361	6649 N Blue Gum St	Schaumburg	Cook	IL	60173	US	I	A	423920	5
20130100002	SHANTI ROBB	789219943	4 B Blue Ridge Blvd	Sacramento	Sacramento	CA	95827	US	I	G	621111	7
20130100003	SCOT CLEMENT	985406053	8 W Cerritos Ave #54	Campbell	Santa Clara	CA	95008	US	I	S	522293	2
20130100004	HARRIET BLYTH	543516468	639 Main St	Waipahu	Honolulu	HI	96797	US	I	B	621991	1
20130100005	DI CRISTINA J & SON	359462084	34 Center St	Brooklyn	Kings	NY	11204	US	C	A	515120	9
20130100006	AMERICAN ARTS & GRAPHICS	760027983	3 McAuley Dr	Springfield	Greene	MO	65803	US	C	Q	423920	15
20130100007	TREENA LARVE	198816592	7 Eads St	El Paso	El Paso	TX	79925	US	I	E	322130	1
20130100008	GUSTAVO HALWOOD	846274720	7 W Jackson Blvd	Somerville	Somerset	NJ	08876	US	I	C	515120	2
20130100009	WILFRIED FITSCHEN	181213075	5 Rector Ave #88	Flushing	Queens	NY	11355	US	I	A	322111	4

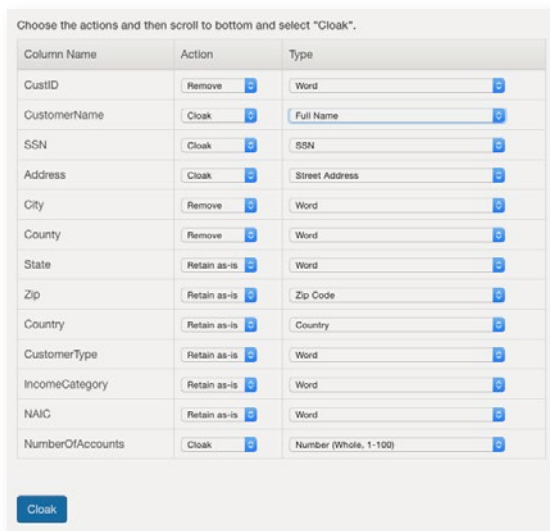
CustID	20130100001
Customer Name	MERRILL MEISENBURG
SSN	688297361
Address	6649 N Blue Gum St
City	Schaumburg
County	Cook
State	IL
Zip	60173
Country	US
Customer Type	I
Income Category	A
NAIC	423920
Number Of Accounts	5

This table shows the names of the data attributes and their values for the first customer record. The names of the data attributes are taken from the first row in the spreadsheet.



THE DATA CLOAK INTERFACE

The screen shot above shows a web browser running the Alacer Data Cloak application. The user selects the “Browse files” action and is prompted to select an input file.



After the user confirms the selection, Data Cloak generates a table (shown left). The table contains the names of the all the first-row headers in the CSV input file. For each header, it offers drop-down lists of choices for “Action” and “Type.” The actions are:

- Remove: Do not include this attribute in the public output
- Cloak: Keep this attribute, but do not reveal its true values. If this action is selected, the “Type” column can be used to control how the sensitive data are cloaked.
- Retain as-is: The data are not sensitive and do not need to be removed or cloaked.

In this example, the customer ID, city, and county will be removed completely. The customer name, social security number (SSN), street address, and the number of accounts will all be cloaked. The remaining attributes will be left as they are in the public data file.

When the user selects the “Cloak” actions, the application generates two files: public file and private file.

Data Cloak output

PUBLIC FILE

The public file generated by the application is intended to be distributed to any outside party. The private file, containing the sensitive information, is intended to be kept by the data owner.

CLOAK_ID	CustID	CustomerName	SSN	Address	State	Zip	Country	Customer Type	Income Category	NAIC	Number Of Accounts
0b06d3c324f350ca37e9e4a037a7aaca66faac17	jobilhumazcu	Lilly Shelton	860629772	314 Tavjo Key	IL	60173	US	I	A	423920	13
890148aadbd5ec6d41a0865bce3cea11a00faeb9	onishajojda	Owen Lloyd	971884632	1227 Tulvo Key	CA	95827	US	I	G	621111	46
d36d61c8ae6268831032f8d41572b8431cbf7eb6	uvolitejtadtin	Lucy Chavez	433760178	512 Sezu Ln	CA	95008	US	I	S	522293	96
453f524360c45217d1f6c7b347312c0cd84853f5	recoabirukes	Gene Stephens	96178757	152 Berne Park	HI	96797	US	I	B	621991	82
da734ff6c8be88fde718166edff5af75392e0f46	wuwezroshoso	Tyler Gordon	882757950	1724 Rodnar Way	NY	11204	US	C	A	515120	1
a2bcb8b50e66a2e9508ba5a77bdd35b73bfa7eca	fikirikfizap	Katie Simon	570934928	1863 Latled Ln	MO	65803	US	C	Q	423920	63
1cd431955ba87a9591e8c2c8d4cb7900e3001a6f	papfevjimoneg	Derek Phillips	588576904	996 Fuci Pt	TX	79925	US	I	E	322130	82

The table above shows the first several rows of the sanitized public file. The public file contains a unique CLOAK_ID identifier for each record along with the retained and cloaked attributes.

Examples below compare the attributes and values of the original file with the sanitized public file. The customer ID, city and county have been removed from the public data. The customer name, SSN, street address, and number of accounts have different, randomly generated values in the public file.

ORIGINAL UNCLOAKED FILE

CustID	20130100001
Customer Name	MERRILL MEISENBURG
SSN	688297361
Address	6649 N Blue Gum St
City	Schaumburg
County	Cook
State	IL
Zip	60173
Country	US
Customer Type	I
Income Category	A
NAIC	423920
Number Of Accounts	5

PUBLIC FILE

CLOAK_ID	0b06d3c324f350ca37e9e4a037a7aaca66faac17
CustID	jobilhumazcu
CustomerName	Lilly Shelton
SSN	860629772
Address	314 Tavjo Key
State	IL
Zip	60173
Country	US
Customer Type	I
Income Category	A
NAIC	423920
Number Of Accounts	13

PRIVATE FILE

The private file contains all of the attributes from the original data, as well as the cloaked data in the public file. This table lists the attributes

CLOAK__ID	0b06d3c324f350ca37e9e4a037a7aaca66faac17
CustID	20130100001
CLOAK__CustID	jobilhumazcu
CustomerName	MERRILL MEISENBURG
CLOAK__CustomerName	Lilly Shelton
SSN	688297361
CLOAK__SSN	860629772
Address	6649 N Blue Gum St
CLOAK__Address	314 Tavjo Key
City	Schaumburg
County	Cook
Number Of Accounts	5
CLOAK__Number Of Accounts	13

and their values for the first record in the private file:

This table shows the first columns and rows of the private file:

CLOAK__ID	CustID	CLOAK__CustID	CustomerName	CLOAK__CustomerName
0b06d3c324f350ca37e9e4a037a7aaca66faac17	20130100001	jobilhumazcu	MERRILL MEISENBURG	Lilly Shelton
890148aadbd5ec6d41a0865bce3cea11a00faeb9	20130100002	onishajojda	SHANTI ROBB	Owen Lloyd
d36d61c8ae6268831032f8d41572b8431cbf7eb6	20130100003	uvolitejtadtin	SCOT CLEMENT	Lucy Chavez
453f524360c45217d1f6c7b347312c0cd84853f5	20130100004	recroabirukes	HARRIET BLYTH	Gene Stephens
da734ff6c8be88fde718166edff5af75392e0f46	20130100005	wuwezroshoso	DI CRISTINA J & SON	Tyler Gordon
a2bcb8b50e66a2e9508ba5a77bdd35b73bfa7eca	20130100006	fikirikfizap	AMERICAN ARTS & GRAPHICS	Katie Simon
1cd431955ba87a9591e8c2c8c4cb7900e3001a6f	20130100007	papfevjimoneg	TREENA LARVE	Derek Phillips
58a1ca7b7d6e1a9bdc8c02ea82da9bb80f080dae	20130100008	moomlohame	GUSTAVO HALWOOD	Gary Foster
49f7c5d0daa38a202445c29788ee239f4765141d	20130100009	gutaregsupe	JUERGEN FITSCHEN	Betty Maldonado

Alacer Data Uncloak application

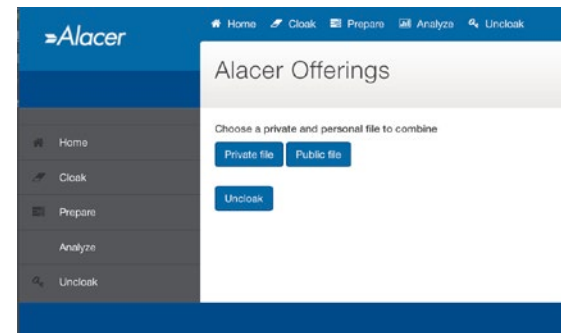
Continuing the money-laundering analysis example, the bank department would like to share the public data file with an outside analyst. The analyst constructs a risk model and assigns each customer record a risk rating value

CustomerName	SSN	Address	State	Zip	Country	Customer Type	Income Category	NAIC	Number Of Accounts	Risk Rating
Lilly Shelton	860629772	314 Tavjo Key	IL	60173	US	I	A	423920	13	8
Owen Lloyd	971884632	1227 Tulvo Key	CA	95827	US	I	G	621111	46	3
Lucy Chavez	433760178	512 Sezu Ln	CA	95008	US	I	S	522293	96	2
Gene Stephens	96178757	152 Berne Park	HI	96797	US	I	B	621991	82	1
Tyler Gordon	882757950	1724 Rodnar Way	NY	11204	US	C	A	515120	1	5

between 1 and 10. This table shows the first few rows of the enhanced public data.

The outside analyst sends the enhanced data file back to the data owner with a risk rating associated with each customer record. For example, customer Eddie Perez has a risk rating of 8. Since this is the public data file, “Eddie Perez” is a randomly generated name. Eddie’s SSN and street address are also randomly generated. How does the data owner recover the customer’s true identify?

The Alacer Uncloak application allows the user to merge matching private and public files into a new file. Like the Data Cloak application, browser-based Alacer Data Uncloak does not transmit data over the network. It creates a single file that merges the enhanced or modified public data with the original private data.



CustID	20130100001
CLOAK__CustID	jobilhumazcu
CustomerName	MERRILL MEISENBURG
CLOAK__CustomerName	Lilly Shelton
SSN	688297361
CLOAK__SSN	860629772
Address	6649 N Blue Gum St
CLOAK__Address	314 Tavjo Key
City	Schaumburg
County	Cook
Number Of Accounts	5
CLOAK__Number Of Accounts	13
State	IL
Zip	60173
Country	US
CustomerType	I
IncomeCategory	A
NAIC	423920
Risk Rating	8

The user selects a public and private file and Uncloak weaves them together to create a new file. The example on the left shows the attribute names and values of the first record in the new output.

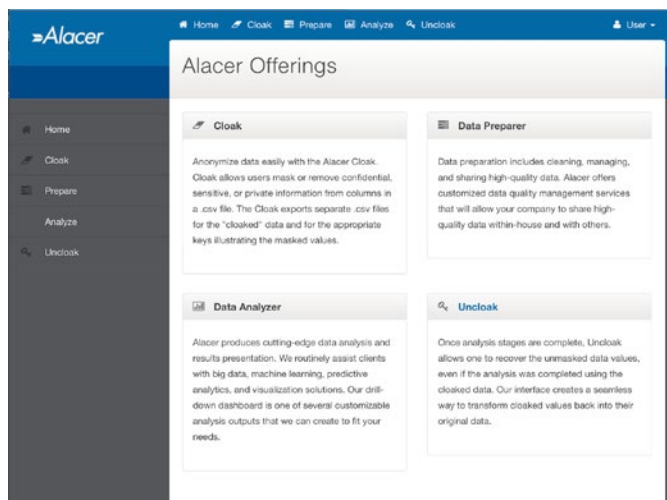
The data owner can now connect the risk rating 8 to the true customer “Merrill Meisenburg.” If additional analysis is required, the cloaked fields are available for easy reference.

Start cloaking your sensitive data now

Alacer's Data Cloak and Uncloak applications are simple, security-tested and affordable. They streamline managing data files for any organization and provide a new standard of protection for sharing sensitive data.

The new Alacer Data Cloak

- Run off-line using existing browser software
- Secure sensitive data with reversible cloaking
- Powerful controls in easy-to-use interface
- Share data freely without risk



TRY THE ALACER DATA CLOAK

Call 800 414-5170 to arrange a demonstration today.

Alacer Data Cloak is patent pending.

Aaron Hoffer is the Alacer Technology Group Senior Solution Architect. He created the Alacer Center for Advanced Computing where he leads development of social network, Internet of Things, business analysis and intelligence reporting tools for financial services clients. Areas of special interest include distributed and concurrent systems, machine learning and all aspects of data science. Prior to Alacer, he held positions at T-Mobile, CIBER, JP Morgan Chase & Co, Saguaro Systems and IBM. Aaron holds a Masters degree in Computer Science and Software Engineering from the University of Washington with a Bachelor of Science degree in Mechanical Engineering.