HUSH HUSH
SENSITIVE DATA DISCOVERY TOOL MANUAL

Table of Contents

INTRODUCTION

Hush Hush Sensitive Discovery Tool is a Windows based desktop utility. Its purpose is to find sensitive data in the database, create workflows to de-identify it and save the metadata for the audit purposes. The tool is used currently with SQL Server and mySQL databases, both on premises and hosted as virtual machines in Azure marketplace, and creates SSIS workflows that use SSIS data masking components to de-identify sensitive data.

SENSITIVE DATA DEFINITION

Sensitive data is data that allows other people to identify you as a person within other records. If someone steals or accesses your sensitive data without permission, s/he could do irreparable harm through credit card fraud, medical records fraud and other forms of identity fraud. There are a lot of laws protecting your identity and data; however, for the purposes of this document we need to concern ourselves with those acts that protect data in development and integration, such as GLBA, HIPAA, PCI/DSS, GDPR, and local countries' state and municipalities initiatives.

Other names for sensitive data include PII (personally identifiable information), PHI (personal health information), private data, direct and indirect identifiers, etc.

The domain of sensitive data and de-identification was introduced by Latanya Sweeny, and she was also the first to define models for sensitive data. In particular, such **attributes as names, addresses, cities, zip codes, dates, VINs, driver licenses, passport numbers, SSN/SIN and other forms of IDs, telephone numbers, emails** all constitute sensitive data. The model includes unique identifiers (**direct identifiers** in another classification) such as SSN: you and only you have your own personal unique SSN. Direct identifiers include information that relates specifically to an individual such as the individual's residence, including for example, name, address, Social Security Number or other identifying number or code, telephone number, e-mail address, or biometric record. It also contains indirect identifiers. Non-unique identifiers (**Indirect identifiers)** include information that can be combined with other information to identify specific individuals, including, for example, a combination of gender, birth date, geographic indicator and other descriptors. Other examples of indirect identifiers include place of birth, race, religion, weight, activities, employment information, medical information, education information, and financial information.

In some industries, the basic sensitive data model has been described in case there is no expertise "in-house". For example, HIPAA lists 18 elements you would need to mask, and calls this model "Safe Harbor". This is by far not the best, but rather sufficient model for de-identification. HIPAA considers expert determination of the model to be the best method for data de-identification. In order to achieve the best results with expert determination, one has to

understand types of attacks and industry related metrics, such as k-anonymity and l-diversity (see Hush-Hush white paper on the topic).

Other industries do not necessarily have "Safe Harbor". Their attributes will also include other set of metadata. However, the methods to determine which data is sensitive and needs to be de-identified are the same.

HOW TOOL HELPS TO DETERMINE SENSITIVE DATA

Hush-Hush Sensitive Data Discovery tool uses Safe Harbor and some other pre-defined elements as a base for the model as well as allows the practitioner to add metadata to the model.

The proprietary algorithm (Patent pending) searches databases' metadata, data patterns and values, and assigns rating to the "suspected" attributes based on presented sensitive data type. Sensitive data types include Name, Last Name, Street Address, City, State, Country, Zip, Phone, Generic Alpha Numeric ID, SSN, SIN, Credit Card, PAN, Driver License, Numeric, Date of Birth, Email, VINs.

The search is not exhaustive, and uses subsamples based on statistical "popularity" of data in USA and Canada, if metadata is not properly named. To use a complete exhaustive search would be impractical in case of large data sets; thus, the tradeoff. [1]

ALGORITHMS SUGGESTIONS

Upon determining sensitive data's fields, the tool suggests the algorithms that it finds the most appropriate. Those fields that have primary-foreign key relationships most likely will be assigned deterministic data de-identification algorithms used by the components.

For a complete list of algorithms, please refer to the SSIS components manual and Wikipedia on http://mask-me.net site.

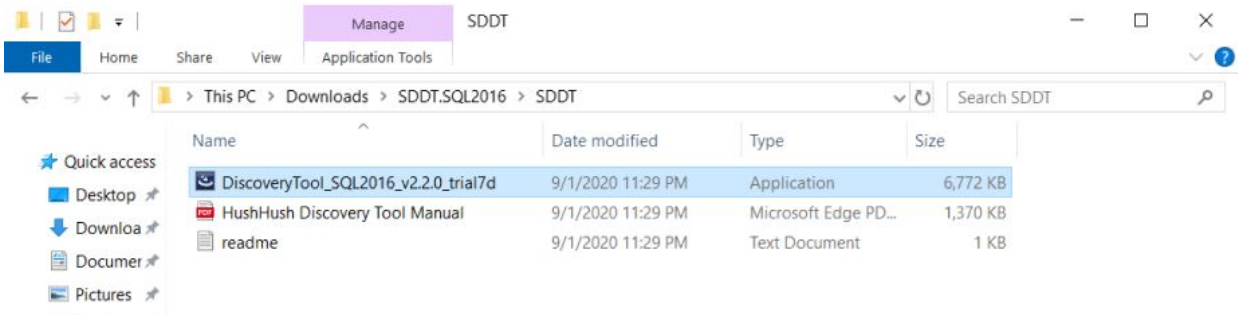
[1] It is recommended to use tool's suggestions as a guideline only and conduct expert determination to complete the determination of the sensitive data set , especially on sparsely populated databases.

STEP BY STEP INSTRUCTIONS

Installation

1. Prerequisites
   1.1. Prerequisites for Sensitive Discovery Tool
   To use Sensitive Discovery Tool you should have:
   Operating System: Windows 8.1, Windows 10 or Windows Server 2016
   Visual Studio 2017 or Visual Studio 2019 with installed SSDT BI
   SQL Server: MS SQL Server 2012, 2014, 2016 with installed SSIS
   Connectors for Oracle and MySQL(Optional)
   Hush-Hush SSIS Components (Required for generation packages)
   .Net Framework v4.6.1 or above

2. Install Hush-Hush Sensitive Discovery Tool
   2.1. After downloading archive unzip it and run installation file

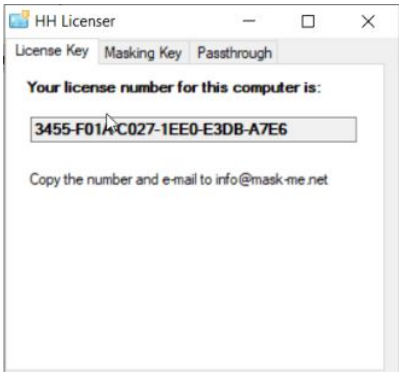


   2.2. In InstallShield wizard window press “Next”
   2.3. Read and accept license agreement and then press “Next”
   2.4. Select destination folder and press “Next”
   2.5. Press “Install”
   2.6. Allow application to make changes on your computer
   2.7. Click “Finish” to finish installation

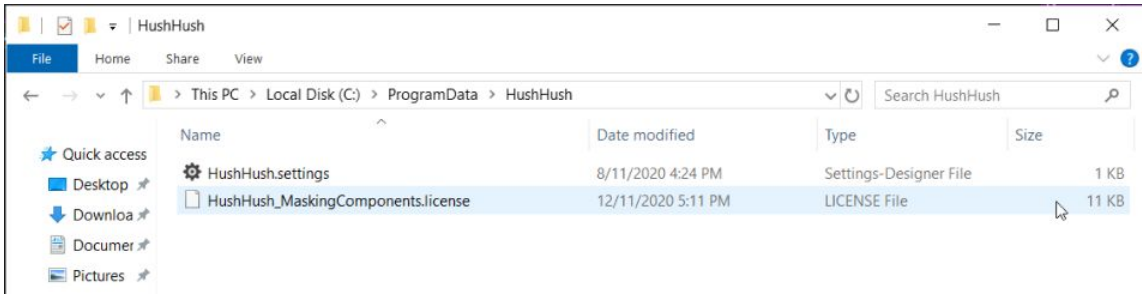
3. To verify that the tool works correctly, use steps in the next section (Connecting to the source and destination) to connect to your database.
4. Install connectors for Oracle database website (Optional)
5. Install connectors for MSSQL database website (Optional)

6 Registering a license (for this step Hush-Hush SSIS Components should be installed)

Our licenses are bound to a specific hardware configuration. To get your hardware ID, open folder C:\Program Files (x86)\HushHush\SSIS\Tools and run “HushHushLicenser.exe”. Copy the generated Hardware ID and send it to info@mask-me.net



After purchasing a license, you will receive a special key file, which should be placed in the following directory: C:\ProgramData\HushHush

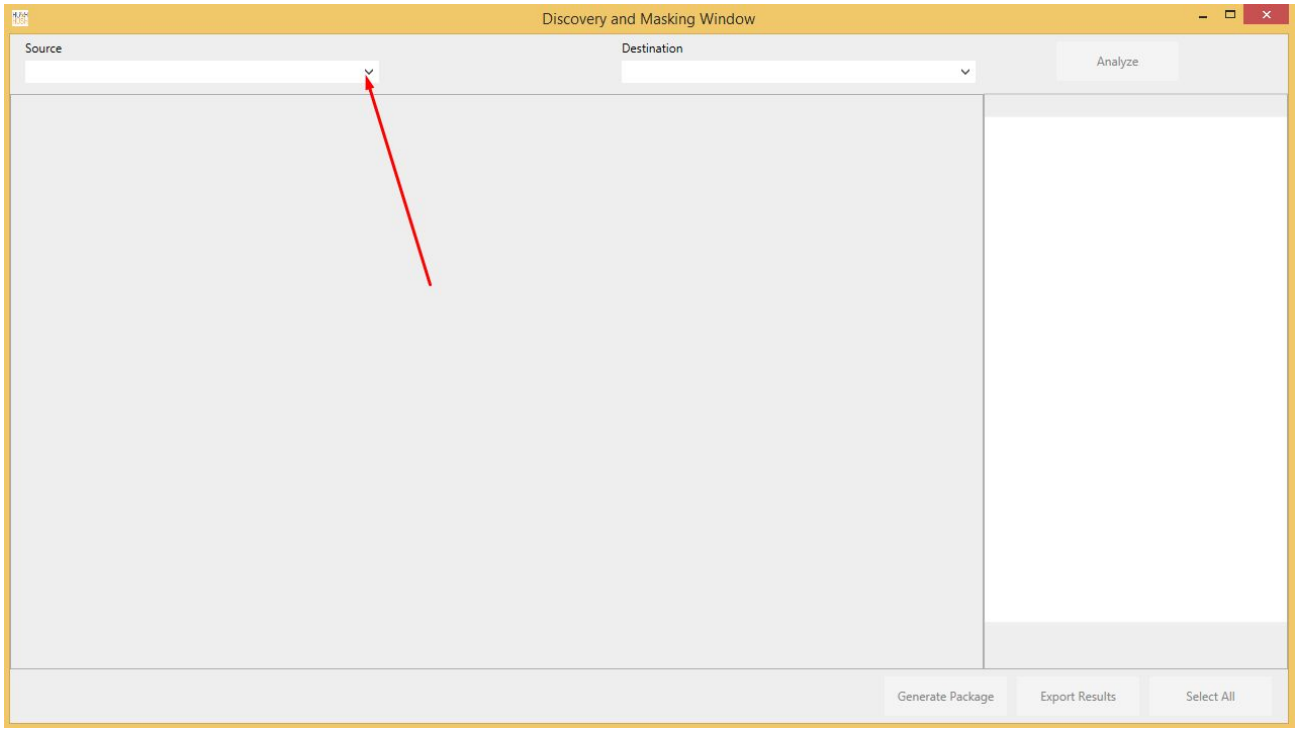


Connecting to the source and destination

First, you would need to connect to both source of your data (production environment tables) and the destination (non-production environment tables). The tool provides standard connection prompts for the OLE DB (when available) or ODBC (when OLE DB is not available) connections.

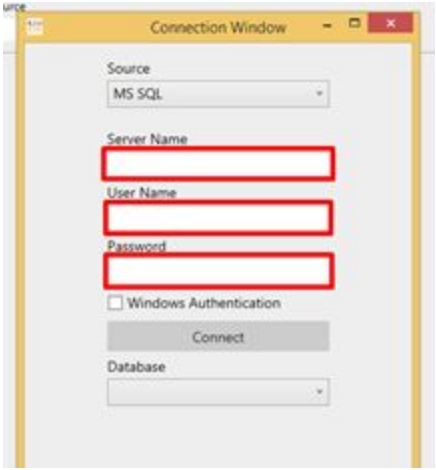
Select Source Connection (Server with initial data)

Click on source dropdown field.



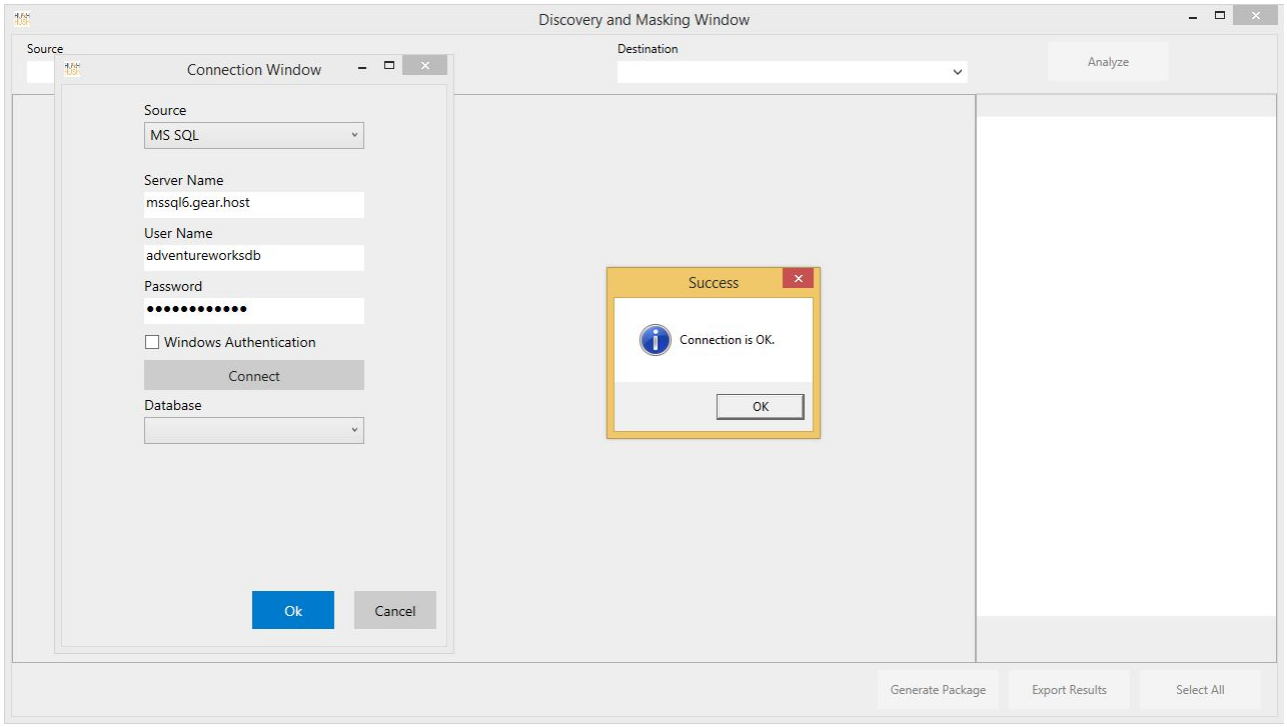
Select «New Connect to Database» option.

In «Connection Window» select the type of the server. Based on the connection driver, you will be prompted for the needed server connection information.

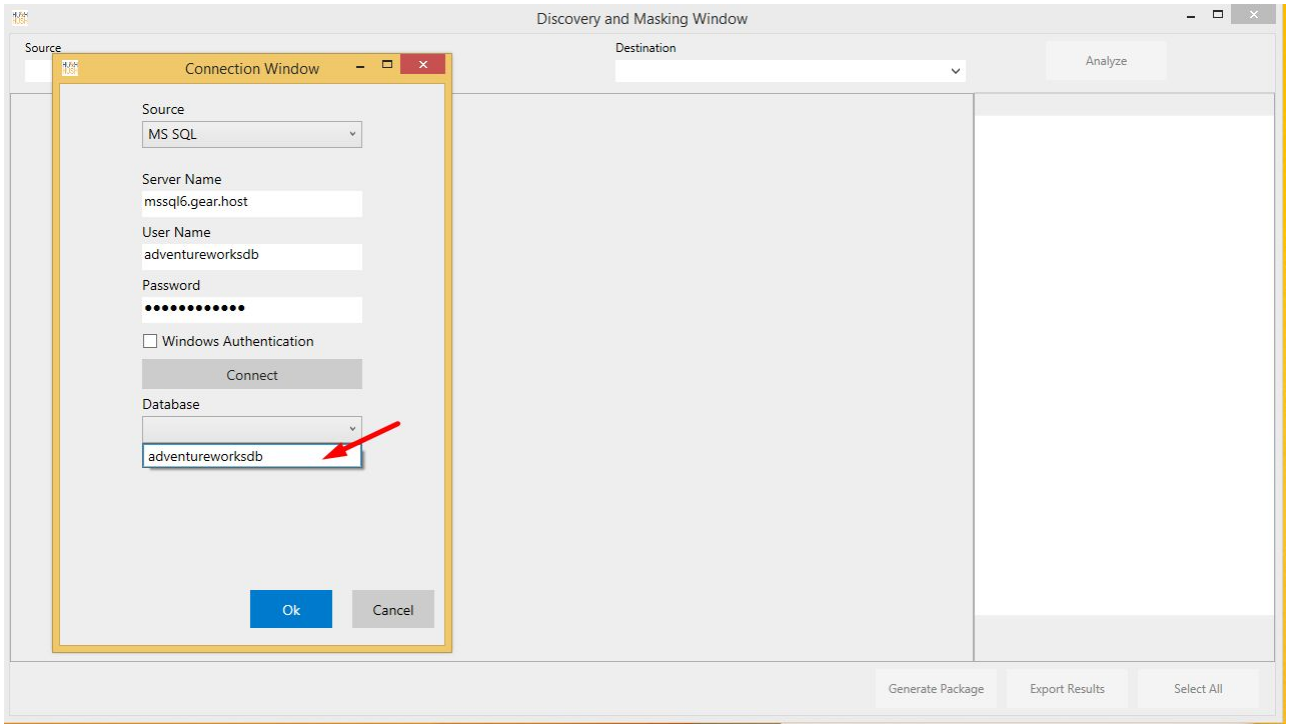


Please, enter your credentials in order to connect.

After you entered your data and pressed «Connect», you should receive the confirmation popup.



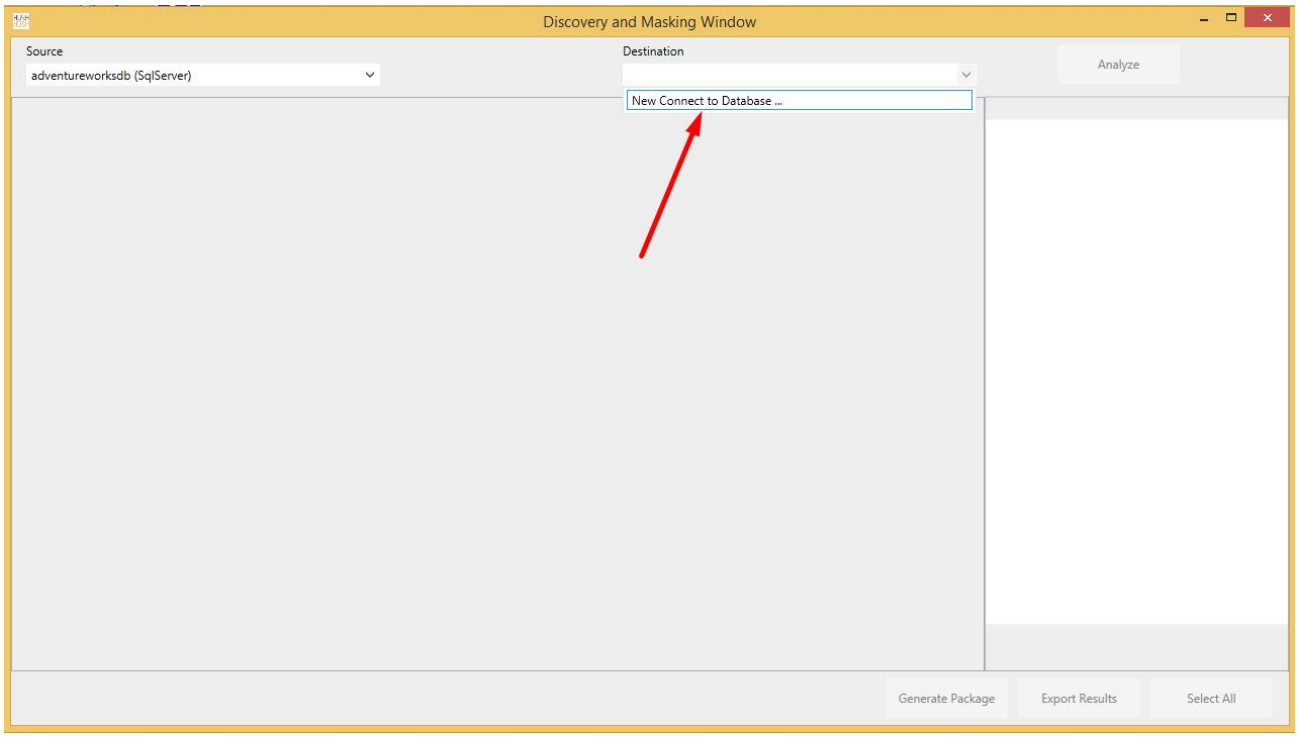
After successful server connection, choose the database from the “Database” dropdown menu. It will contain the list of the databases accessible per provided login.



Click «Ok»

Now, select a Destination connection (Server for output from discovery tool)

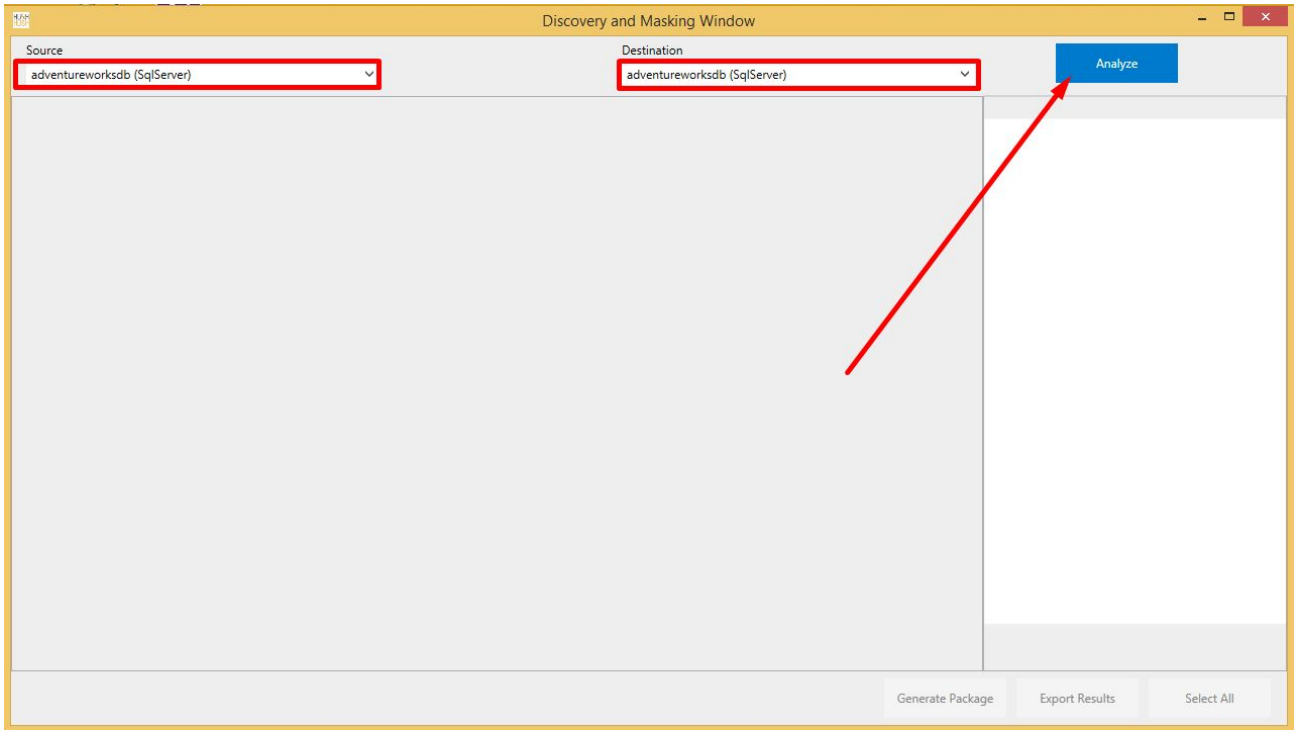
Click on the Destination dropdown field.



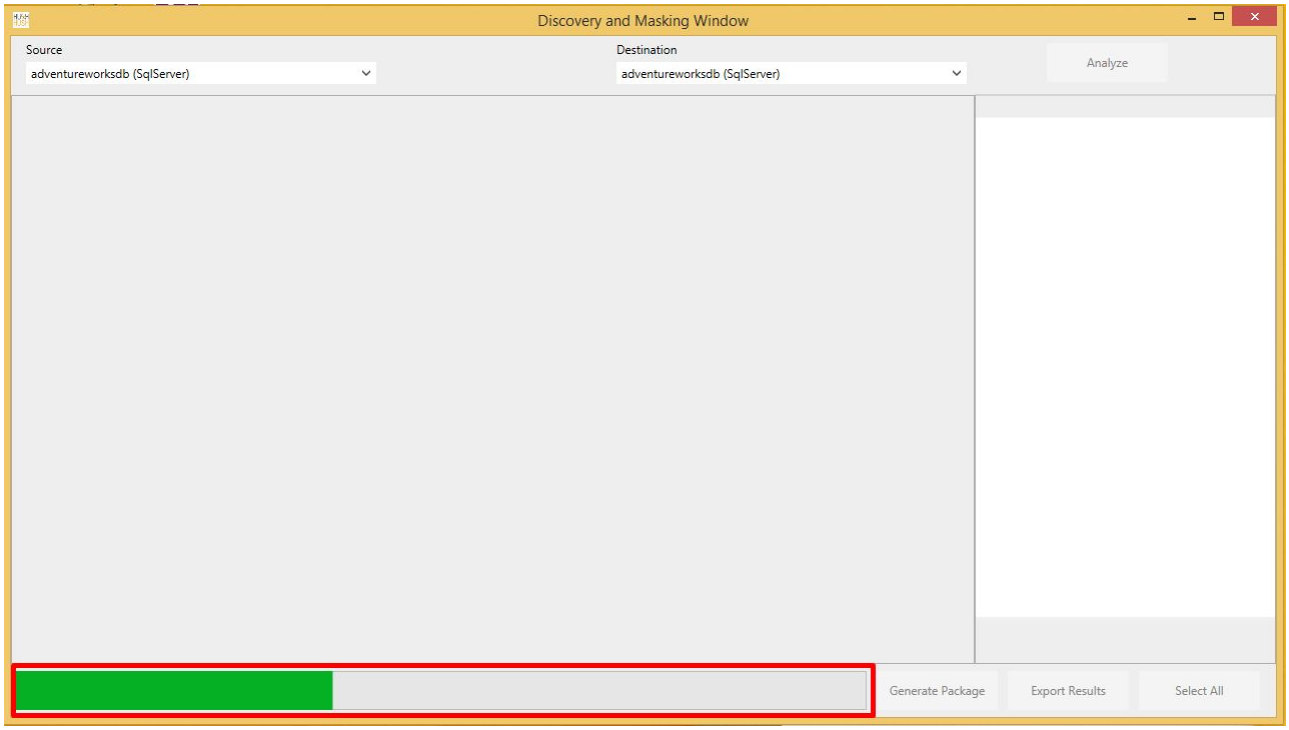
In «Connection Window», select the type of the server and enter your credentials.

Analyzing databases for the sensitive data determination

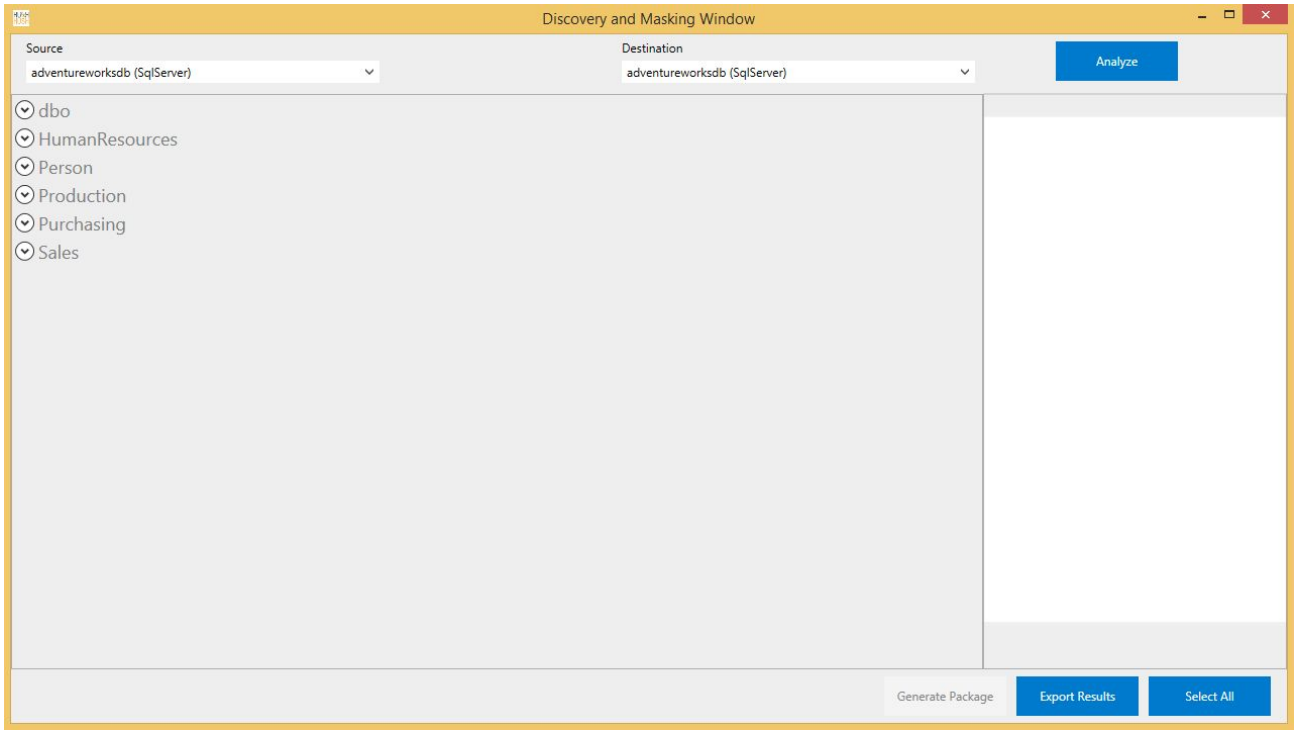
After you have connected to both Source and Destination servers, click on “Analyze” button in the right upper corner.



The system will go against current database metadata and data and apply search patterns. Green bar below will display current progress of the discovery process.



After analysis is done, the tool displays the database schema information and identifies which one is pertaining to the sensitive data. The metadata display format is <schema>.<table>.<field>. Below you can see a list of schemas in our sample database AdventureWorksDB. Clicking on the schema arrow will expand it, and you will see multiple tables with sensitive data and suggested algorithms.

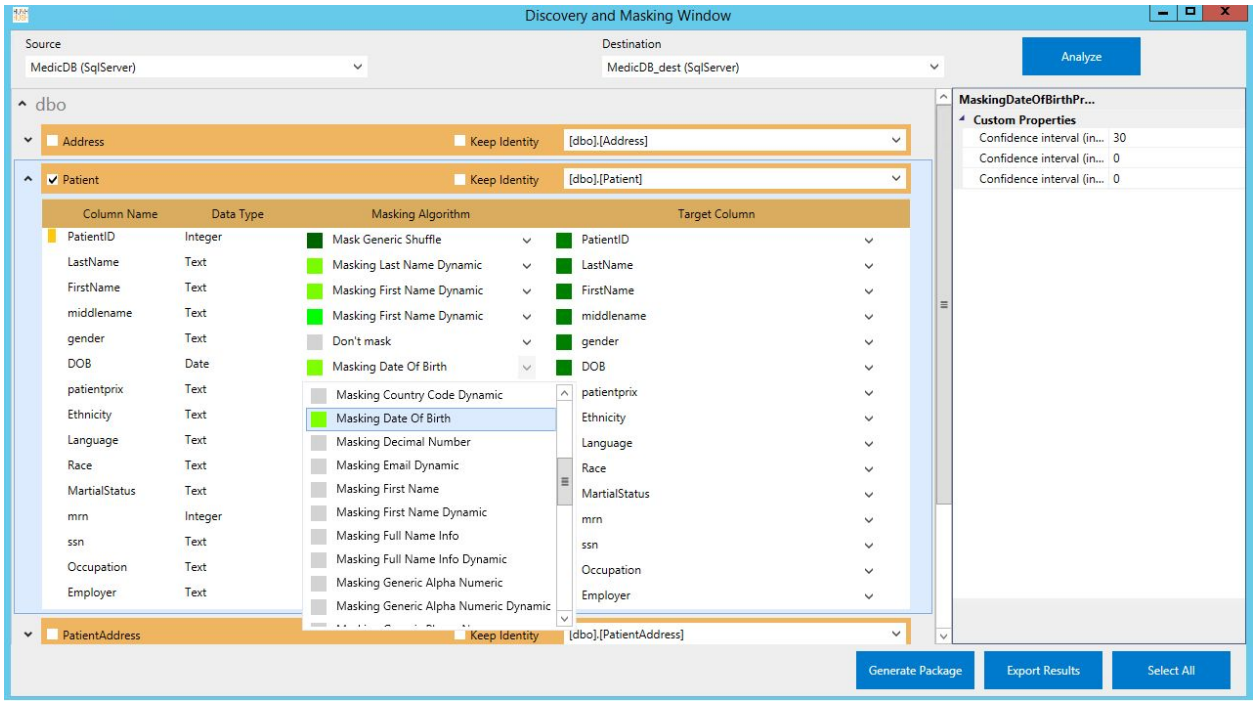


If you will open any table after analysis process, you will see that the tool determined the type of data and selected the closest algorithm based on metadata names and values for data masking. The search for data types is based on HIPAA's Safe Harbor. However, it does not accommodate Expert Determination, as it is limitless. We provide you with the capability to choose algorithms manually.

Expert Determination – supplementing Safe Harbor

Masking algorithm column in the tool presents multiple drop downs, containing all the available algorithms, both specific and generic, that are used in the Hush-Hush data masking components. Generic algorithms available via tool include Shuffle, number, date and string RollUps, and Generic Alpha Numeric algorithms. Generic algorithms take any value as an input. Other algorithms are sensitive data type specific and come in several varieties: there are dictionary based algorithms for indirect identifiers, and pattern/ISO based algorithms for direct identifiers such as SSNs, Credit Cards, etc. If there is no pattern based or ISO algorithm available – you can always use generic one. Dictionary substitution is not currently used and will be introduced in the future versions of the tool. For more specific use of the algorithms, please refer to the SSIS components manual and to the Wikipedia on the http://mask-me.net site.

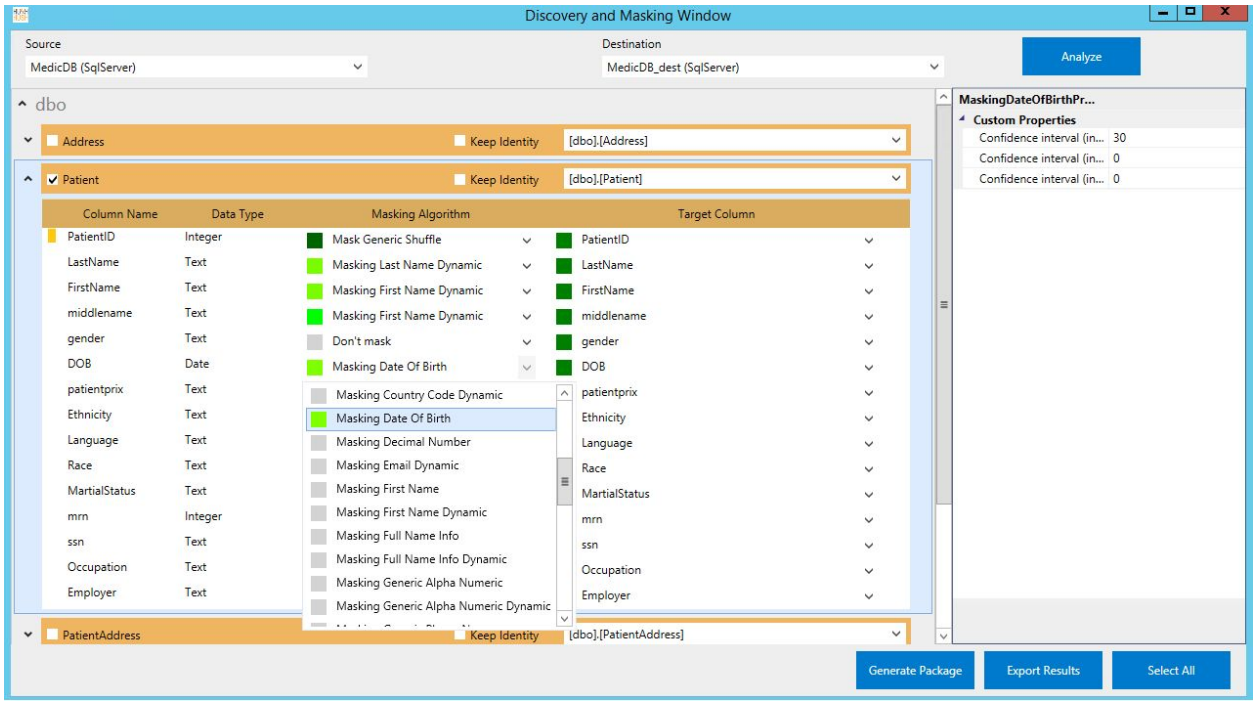
At this step of the process, you would review suggested algorithms and make proper selection to adjust if necessary.



Also, you can pick output table and output column for further data load.

Color next to the algorithm indicate the probability of proper data type/algorithm match and corresponds to the probability index. You see black if the components are not installed.

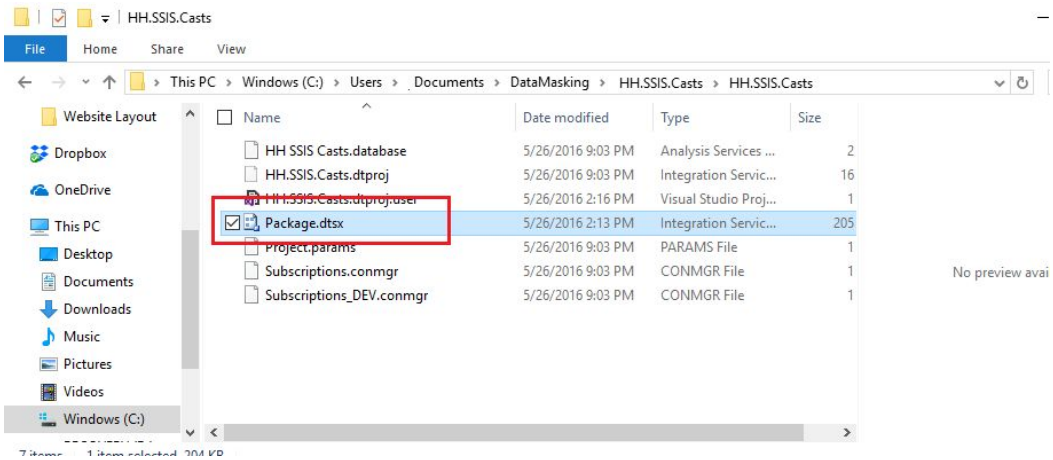
You will see different colors of green if the components are installed on the machine that has Sensitive Data Discovery utility installed.



Some components require configuration with the default and parameter values. Clicking on the algorithm will populate the Properties window per algorithm/component, and you can make necessary data entry.

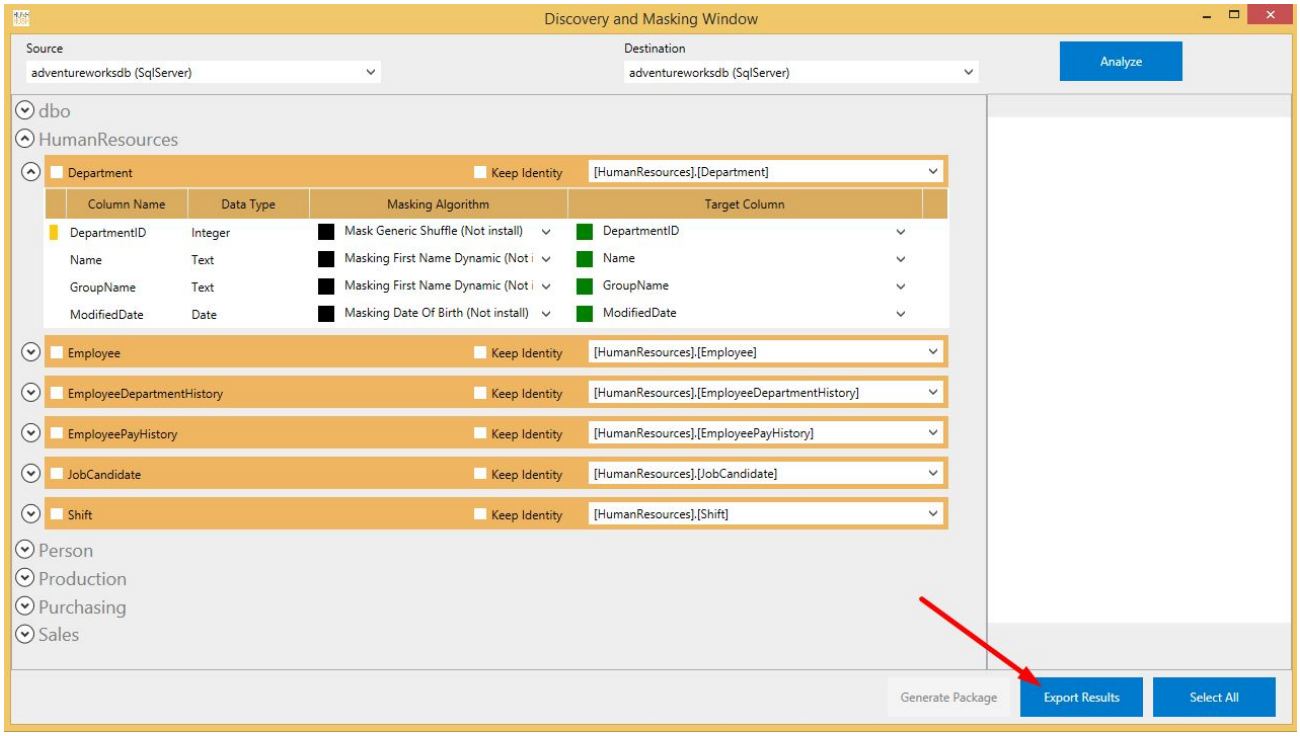
Generating Packages

The tool's ultimate purpose is to generate SSIS package with data masking workflow. In order to do that, select the checkbox near table name, or click "Select All" to choose all tables. Than click "Generate Package" button. The tool generates an SSIS package and prompts you to save on disk. When you save a package in the directory, it is your typical SSIS package:

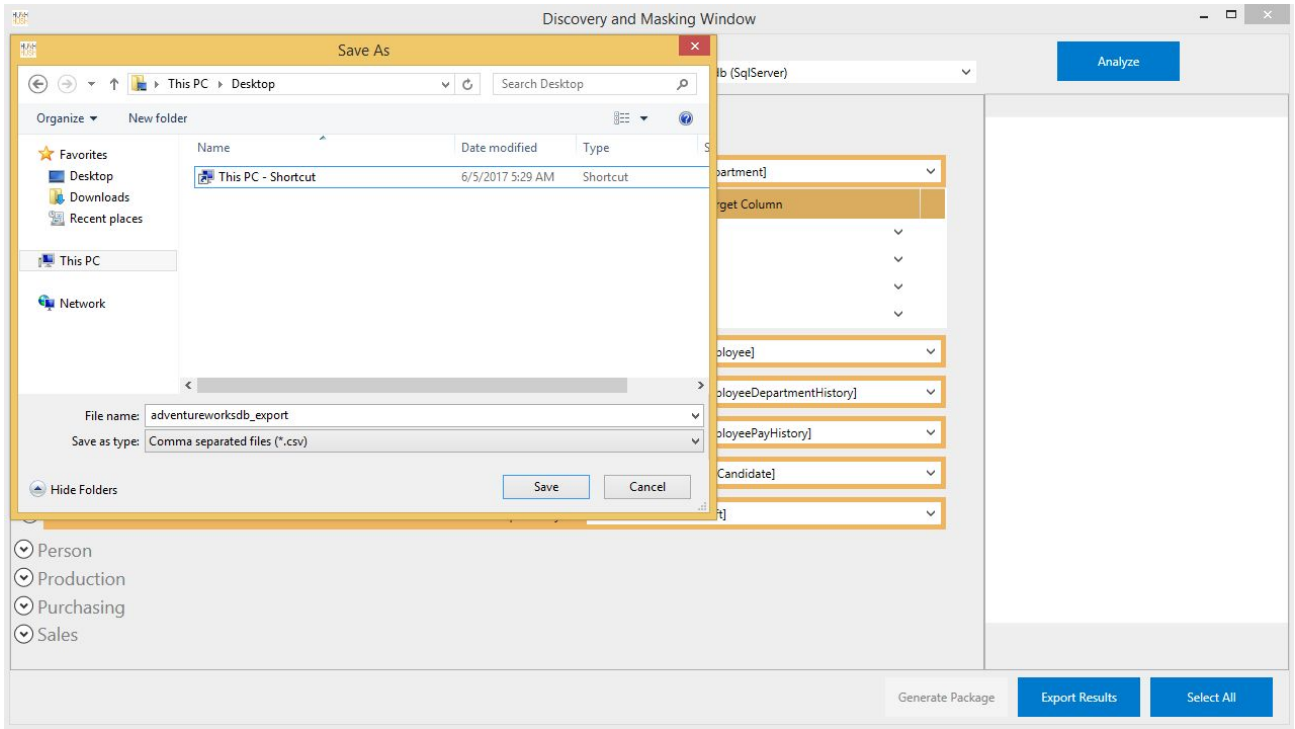


Exporting metadata for Audit

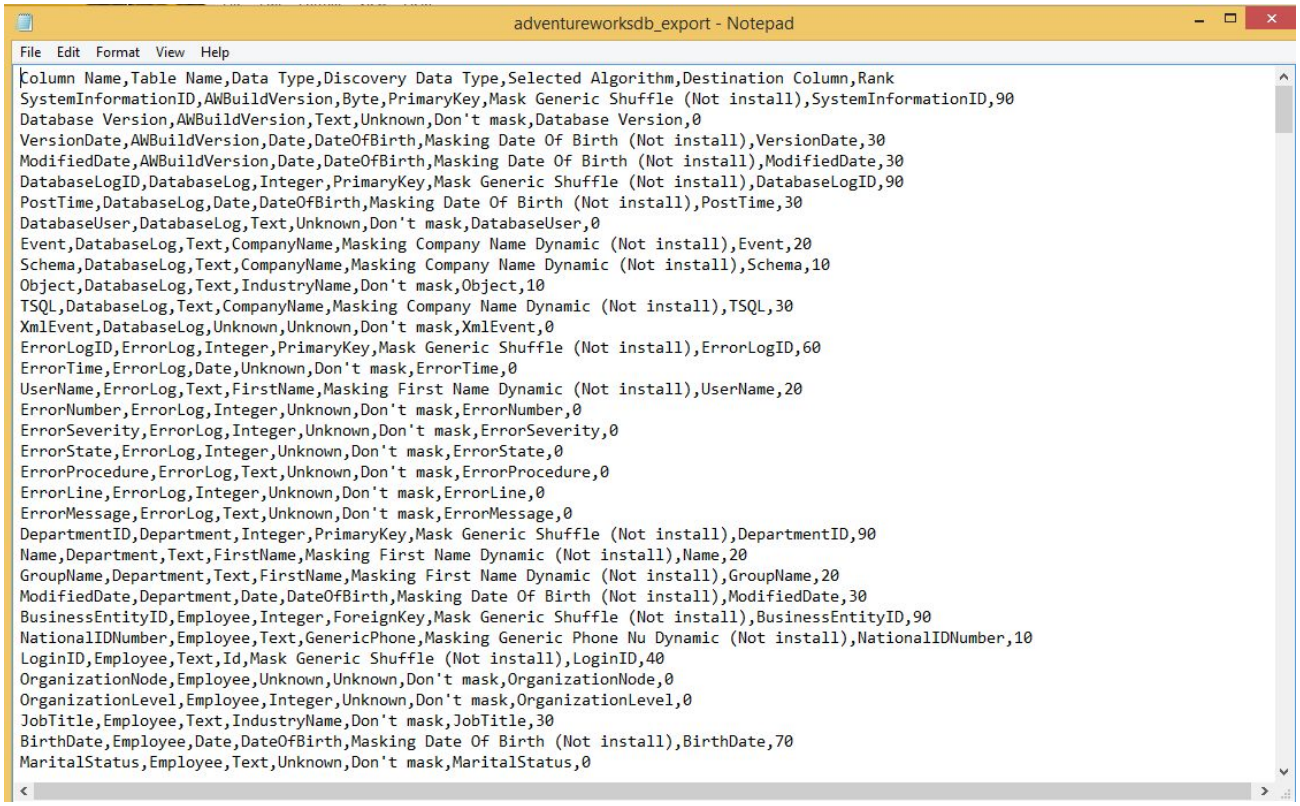
Even if you do not use the components, but want to discover Sensitive data, and save for future use, you can utilize the tool. Please, notice that if the components are not installed, the button "Generate Package" is not enabled. Clicking on the button "Export Results" uploads the results of search on the current database into the csv file.



You can select the path to save your file:



The screenshot below shows the result output in the file. The file includes rank field that has quantative probability of determining the algorithm of the column based on patent-pending method.



INTRODUCED IN THIS RELEASE:

This is the first release.

COMING SOON

Support for Oracle, and Generic Dictionary Substitution component.

CUSTOM COMPONENTS

Please contact us if there is a custom component that you need **in your particular situation!**

TABLE OF COMPONENTS AND ALGORITHMS

component	*random*	*dynamic*	*date and number variance*
SSN	X	X	
Credit Card	X	X	
First Name	X	X	
Last Name	X	X	
Address	X	X	
City Names	X	X	
US Phone Number	X	X	
Email	X		
ZIP	X	X	
Date of Birth			X
URL	X	X	
Generic Alpha Numeric	X	X	
Phone Number	X	X	
Country Codes	X	X	
Decimal Number			X
State Province	X	X	
Dictionary Load – generic			
Name with Gender	x	x	
SIN	x	x	
Ip Address	X	x	
Full Name	X	X	
Shuffle – Generic			
Substitution - Generic			
Company Name	X	X	
Number Rollup		X	
String Rollup		X	
Date Rollup		x	