Data Formats
HashSets.com Databases

Within this updated document we have added new changes to the number of tables and their names and small additions to the table structures. There is no removal of data fields within this update. All changes will be highlighted in Blue for faster reference. Please note the changes below and implement into your own preexisting or newly created database and tables.

Updated: 23 November 20243

1. Introduction

This document describes the formats of various database tables provided by HashSets.com to customers who subscribe to our 'Database Exports' (also known as 'database dumps'). Throughout this document we will cover only trusted, known-good and/or non-threatening hash set tables from our database.

Since November 2003, Whitehat Computer Forensics, LLC, has been performing hash file calculations of common electronic files found within various types of computer operating systems, workstations and servers.

Our largest hash set consist of operating system files from more than 700+ operating system versions we have painstakingly installed onto either sterile hard disks and/or virtual drives, subsequently applied the latest manufacturer updates when available, and forensically analyzed using commercially acceptable computer forensic software. This allows us to discover and gather all files, file attributes and meta-data and most importantly the MD5, SHA1, SHA-256 and SHA-512 file hash values.

All of the above information is gathered and exported into individual operating system database tables which we use to create our Hash Sets from and place into our online search engine(s). We concurrently offer to Gold and Platinum subscription members our database tables which are exported into large tab delimited text files which may then be reimported into their own databases or third-party software products. Each database table also provides a header row for simple field name explanation and identification (see further below for details).

2. Overview of the 'Primary Database Export' Table

The HashSets Database consists of many very large database tables. From these tables there is one master table named '**Primary Database Export**' table.

The '**Primary Database Export**' table contains the permanent computer file information so long as the file is not changed or manipulated by an outside source (e.g. computer users, operating system, software, etc). Permanent file information includes the MD5, SHA1, SHA-256 and SHA-512 hash values, the file's first 32 bytes in hexadecimal format, the first 128 ASCII characters (like using Unix or Linux 'strings' command), logical byte size, and so on.

The '**Primary Database Export**' table contains hash values of only safe, non-threatening, known-good computer files that originated from operating system manufacturers, software and hardware developers, etc, and which were never intended to be used for abuse or malicious purposes. There are also no duplicate hash values within this table. Please remember this table is considered the master table of which all other tables are connected to it.

'Primary Database Export' Table Schema

Field Name	Type	Length	Description
Date_Added	date	0	(Work in Progress) Date that we added the file's hash values and metadata into the 'Primary Database Export' table. Using the following Date format (DD/MM/YYYY). An example would be (20/11/2020)
MD5	char	32	128-bit Message Digest 5 (hash value) of a specific file.
SHA_1	char	40	160-bit Secure Hash Algorithm message digest (hash value) of a specific file.
SHA_256	char	64	256-bit Secure Hash Algorithm message digest (hash value) of a specific file.
Header_HEX	varchar	64	First 32 bytes in Hexadecimal format of the file.
128_Bytes_ASCII	varchar	128	First detectable 128 bytes in ASCII format (Similar to performing Linux or UNIX STRINGS of a file).
Signature	varchar	255	Potential file header signature(s), if known.
Logical_Size	bigint	17	File size in byte format.
NSRL	char	3	File known to be found within the National Software Reference Library (NSRL) Dataset released by the US Government. Will be marked with either 'Yes' or 'No'.
key_field	int	11	Table key that uniquely defines a record, if provided.

SHA_512	char	128	512-bit Secure Hash Algorithm message digest (hash value) of a specific file, if available.
CRC32	char	8	32-bit Checksum of a specific file, if known.
Fuzzy_SSdeep	varchar	255	Context Triggered Piecewise Hash values (CTPH). Also called "fuzzy" hash values. (Future Use Only/Work in Progress).
Ignorable	char	7	Intended to identify potentially ignorable or non-intrinsic hash values. For example, common operating system log files (.log, .evt, .evtx) which may have different contents and hash values if found later on other installations of operating systems, different computers or devices. Will be marked with either 'Yes' or 'Unknown'.
Operating_System_File	char	7	Hash values found within an Operating System. Will be marked with either 'Yes' or 'Unknown'.
Other_Computer_File	char	7	Hash values found outside of operating system installations. For example, third-party software, applications, drivers, utilities, etc, that we downloaded directly from manufacturer websites, etc. Will be marked with either 'Yes' or 'Unknown'.
MS_Windows_OS	char	7	Hash values found within Microsoft Windows operating system installations and updates. Will be marked with either 'Yes' or 'Unknown'.

Linux_OS	char	7	Hash values found within Linux operating system installations and updates. Will be marked with either 'Yes' or 'Unknown'.
BSD_OS	char	7	Hash values found within BSD (UNIX like operating system) installations and updates. Will be marked with either 'Yes' or 'Unknown'.
macOS	char	7	Hash values found within Apple's macOS (formerly OS X) operating system installations and updates. Will be marked with either 'Yes' or 'Unknown'.
Solaris_OS	char	7	Hash values found within Oracle's (formerly Sun) Solaris operating system installations and updates. Will be marked with either 'Yes' or 'Unknown'.
Gold_Disks	char	7	Hash values found within operating system installations with added third-party applications or software deemed to be common use and safe/non-threatening/known-good. These 'Gold Disks' are what we created on our own as a sampling of what other businesses may implement within their own computer networks or environments. Will be marked with either 'Yes' or 'Unknown'.
Installation_Discs	Char	7	Hash values found within various operating system installation media (CD, DVD and .ISO) files. Will be marked with either 'Yes' or 'Unknown'.

3. Other Database Tables (Operating Systems)

The operating system database tables are divided into specific operating systems (MS Windows x86, x64 and Arm64; Linux x86, x64 and Arm64; BSD, macOS and legacy Solaris).

Below is a description of each of the other remaining tables. It should be mentioned briefly that over the 20+ years we noticed the size of our MS Windows, macOS and Linux operating system tables grew too large to contain within single tables each. Therefore, in July 2021 and subsequently in November 2024, we decided to divide those tables into smaller tables for ease of management, transportation and importation by both us and our subscription customers.

MS Windows – These tables were ultimately divided into worldwide geographies (Windows North America, Windows Europe, Windows Asia, Windows Middle-East, etc) and then beginning in late November 2024 subsets of those were created (e.g. x86, x64 and Arm64). Below is their descriptions:

- **North_America_Windows_11_x64_operating_systems** – This database table contains various Microsoft Windows 11 operating system versions for the United States, Canada (French Canadian), Mexico (Spanish Mexican) and the individual file or folder names, file name extensions, file paths, last Modified/Accessed/Created file dates, and so on.

- **North_America_Windows_10_x86_x64_operating_systems** – This database table contains various Microsoft Windows 10 operating system versions for the United States, Canada (French Canadian), Mexico (Spanish Mexican) and the individual file or folder names, file name extensions, file paths, last Modified/Accessed/Created file dates, and so on.

- **North_America_Windows_11_Arm64_operating_systems** – This database table contains various Microsoft Windows 11 Arm64 operating system versions for the United States, Canada (French Canadian), Mexico (Spanish Mexican) and the individual file or folder names, file name extensions, file paths, last Modified/Accessed/Created file dates, and so on.

- **North_America_Windows_Servers_x86_x64_operating_systems** – This database table contains various Microsoft Windows Server operating system versions for the United States (English) and the individual file or folder names, file name extensions, file paths, last Modified/Accessed/Created file dates, and so on.

- **North_America_Windows_Servers_Legacy_x86_x64_operating_systems** – This database table contains various legacy Microsoft Windows Server operating system versions (e.g. MS Windows 2000 Servers, 2003 Servers, 2008 Servers, 2011 Home Servers, 2012 Servers, etc) that are no longer supported by Microsoft. These focused on United States (English) versions and the individual file or folder names, file name extensions, file paths, last Modified/Accessed/Created file dates, and so on.

- **North_America_Windows_Legacy_x86_x64_operating_systems** – This database table contains various legacy Microsoft Windows operating system versions (e.g. MS Windows 2000 Professional, Windows ME, Windows XP, Windows Vista, Windows 7, Windows 8, etc) that are no longer supported by Microsoft. These focused on United States (English) versions and the individual file or folder names, file name extensions, file paths, last Modified/Accessed/Created file dates, and so on.

- **Europe_Windows_10_x86_x64_operating_systems** – This database table contains various Microsoft Windows 10 operating system versions for Europe (Western and Eastern) and the individual file or folder names, file name extensions, file paths, last Modified/Accessed/Created file dates, and so on.

- **Europe_Windows_11_x86_x64_operating_systems** – This database table contains various Microsoft Windows 11 operating system versions for Europe (Western and Eastern) and the individual file or folder names, file name extensions, file paths, last Modified/Accessed/Created file dates, and so on.

- **Europe_Windows_11_Arm64_operating_systems** – This database table contains various Microsoft Windows 11 Arm64 operating system for Europe (Western and Eastern) and the individual file or folder names, file name extensions, file paths, last Modified/Accessed/Created file dates, and so on.

- **Europe_Windows_Servers_x86_x64_operating_systems** – This database table contains various Microsoft Windows Server operating system versions for the United States (English) and the individual file or folder names, file name extensions, file paths, last Modified/Accessed/Created file dates, and so on.

- **Europe_Windows_Legacy_x86_x64_operating_systems** – This database table contains various legacy Microsoft Windows operating system versions (e.g. Windows 8, Windows 8.1, etc) for Europe, if available, that are no longer supported by Microsoft. These focused on United States (English) versions and the individual file or folder names, file name extensions, file paths, last Modified/Accessed/Created file dates, and so on.

- **Asia_Windows_11_x64_operating_systems** – This database table contains various Microsoft Windows 11 operating system versions for Asian countries and the individual file or folder names, file name extensions, file paths, last Modified/Accessed/Created file dates, and so on.

- **Asia_Windows_10_x86_x64_operating_systems** – This database table contains various Microsoft Windows 10 operating system versions for Asian countries and the individual file or folder names, file name extensions, file paths, last Modified/Accessed/Created file dates, and so on.

- **Asia_Windows_11_Arm64_operating_systems** – This database table contains various Microsoft Windows 11 Arm64 operating system versions for Asian countries and the individual file or folder names, file name extensions, file paths, last Modified/Accessed/Created file dates, and so on. ***Note: This is a work in progress effort and will be available sometime in 2025.***

- **Asia_Windows_Servers_x86_x64_operating_systems** – This database table contains various Microsoft Windows Server operating system versions for Asian countries and the individual file or folder names, file name extensions, file paths, last Modified/Accessed/Created file dates, and so on.

- **Middle_East_Windows_11_x64_operating_systems** – This database table contains various Microsoft Windows 11 operating system versions for middle eastern countries or languages and

the individual file or folder names, file name extensions, file paths, last Modified/Accessed/Created file dates, and so on.

- **Middle_East_Windows_10_x86_x64_operating_systems** – This database table contains various Microsoft Windows 10 operating system versions for middle eastern countries or languages and the individual file or folder names, file name extensions, file paths, last Modified/Accessed/Created file dates, and so on.

- **Middle_East_Windows_11_Arm64_operating_systems** – This database table contains various Microsoft Windows 11 Arm64 operating system versions for middle eastern countries or languages and the individual file or folder names, file name extensions, file paths, last Modified/Accessed/Created file dates, and so on. ***Note: This is a work in progress effort and will be available sometime in 2025.***

- **Linux_x86_x64_operating_systems** – This database table contains various Linux operating system distributions and the individual file or folder names, file name extensions, file paths, last Modified/Accessed/Created file dates, and so on.

- **Linux_Arm64_operating_systems** – This database table contains various Linux operating system distributions and the individual file or folder names, file name extensions, file paths, last Modified/Accessed/Created file dates, and so on. ***Note: This is a work in progress effort and will be available sometime in 2025.***

- **Linux_Legacy_x86_x64_operating_systems** – This database table contains various Linux legacy operating system distributions and the individual file or folder names, file name extensions, file paths, last Modified/Accessed/Created file dates, and so on. ***Note: This is a work in progress effort and will be available sometime in 2025.***

macOS – These tables were ultimately divided into two tables (macOS versions 8 thru 10 and macOS versions 11 and above). Below is their description:

- **macOS_8_thru_10_Database_Export** – This database table contains various Apple macOS 8 and 9 (legacy) and macOS 10 (OS X) operating system versions and the individual file or folder names, file name extensions, file paths, last Modified/Accessed/Created file dates, and so on.

- **macOS_11_and_above_Database_Export** – This database table contains Apple macOS 11 and above operating system versions and the individual file or folder names, file name extensions, file paths, last Modified/Accessed/Created file dates, and so on.

BSD – Currently, this is a single table for x86 and x64. In the future we may add Arm64 which will be its own table. Below is its description:

- **BSD_x86_x64_operating_systems** – This database table contains various BSD (UNIX like) operating system distributions and the individual file or folder names, file name extensions, file paths, last Modified/Accessed/Created file dates, and so on. Basically, the individual file or folder information that may not remain the same throughout the purpose of a MS Windows

operating system regardless of individual file hash values (MD5, SHA1, SHA-256 and SHA-512) that generally remain the same until the file itself is ultimately changed or modified.

Solaris – Currently, this is a single table for the Solaris operating system which is considered a legacy operating system. It is most unlikely we will be adding any further data into this table. Below is its description:

- **Solaris_legacy_operating_system** – This database table contains various legacy Solaris operating system versions and the individual file or folder names, file name extensions, file paths, last Modified/Accessed/Created file dates, and so on.

4. Other Database Tables (Non-Operating Systems)

- **Applications_and_Hardware_Software_Drivers_Database_Export** – This database table contains the details of files gathered from third-party utilities and software applications as well as software drivers from common hardware manufacturer websites. Due to the large number of files within this category the installation process (installing onto any ordinary running operating system) before hashing could not be performed. As an alternative, the analysis consisted of unpacking executable files and compressed files, when possible, then subsequently gathering hash value calculations and file meta-data details.

- **Windows_App_Store_Database_Export** – This database table contains MS Windows 8 and 10 Applications commonly found within the MS Windows App Store. Specifically, downloadable business, game, education, and other apps which were installed, analyzed and then gathered into MD5, SHA-1 and SHA-256 hash sets.

- **Mac_App_Store_Database_Export** – This database table contains macOS applications commonly found within the macOS App Store. Specifically, downloadable business, game, education and other apps which were installed, analyzed, hashed and then gathered into MD5, SHA-1 and SHA-256 hash sets.

- **Gold_Disks_Windows_Database_Export** – This database table contains installations of various Microsoft Windows operating systems with the added installations of commonly used programs and software which we have analyzed and hashed. For example, Microsoft Visual Studio, Microsoft Office Suites, common web browsers, etc. Efforts are made periodically to update the Gold Disks installations, but it is best efforts only currently.

- **Microsoft_Windows_Installation_Media_Database_Export** – This database table contains Microsoft Windows installation media (CD, DVD and ISO) which we have gathered, analyzed, hashed and then subsequently placed into MD5, SHA-1, SHA-256 hash sets.

Other Optional Table(s):

- **US_Government_Database_Export** – This optional database table contains common non-threatening known hash values consisting of US Government (federal, state, local and military)

publicly accessible website images, logos, multimedia files, office documents (.doc, .pdf, .xls, .ppt, etc). Please note that this table may be removed completely sometime in the future as these formerly common hash values may no longer be of use or popular to computer forensic or computer security professionals. For the time being we have furnished this table for optional download and use.

- **File_Extensions** – This optional table contains general information or descriptions pertaining various file name extensions (.exe, .pdf, .dll, etc).

  For example, the below file extension .EXE within this table could be described as one of the following from historical popular use:

  EXECUTABLE FILE :::: SELF-DISPLAYING IMAGE :::: SELF-EXTRACTING ARCHIVE :::: SETTLERS 4 SAVE FILE :::: PDP-10 PAGE-MAPPED EXECUTABLE BINARY FILE :::: PLAYSTATION EXECUTABLE FILE :::: OUT-OF-PROCESS CODE COMPONENT FILE :::: MICROSOFT LINKER EXE INPUT FILE EXTENSION :::: DATAFLEX RUNTIME FILE EXTENSION :::: SELF-EXTRACTING ARCHIVE :::: MIME: APPLICATION/OCTET-STREAM FILE EXTENSION :::: MIME: APPLICATION/X-MSDOWNLOAD

  This information above and within the table is used only as one of many starting points when analyzing computer files. It is not to be used as a guarantee that a particular file with a specific file extension is truly associated with any software, program, third-party utility, hardware device, etc.

5. Operating Systems and Non-operating System Table Schema

The following depicts the data elements for the previously mentioned fourteen tables. To associate the 'Primary Database Export' table with any of the above fourteen tables you would use the MD5 hash value from both tables as the linking "Key".

Field Name	Type	Length	Description
MD5	char	32	128-bit Message Digest 5 (hash value) of a specific file.
Name	varchar	255	Names of files and folders.
File_Ext	varchar	255	The file name's extension, if applicable.
Description	varchar	75	A general description of the file or folder.
Last_Accessed	datetime	0	Date and Time that the file was

			last accessed. Uses the following Date and Time format (DD/MM/YYYY Hour:Minute:Seconds). An example would be (20/11/2020 14:06:49)
File_Created	datetime	0	Date and Time that the file was created. Uses the following Date and Time format: DD/MM/YYYY Hour:Minute:Seconds An example would be (20/11/2020 14:06:49)
Last_Written	datetime	0	Date and Time that the file was last modified or written. Uses the following Date and Time format: DD/MM/YYYY Hour:Minute:Seconds An example would be (20/11/2020 14:06:49)
Full_Path	text	0	The full path to the file or folder.
Quick_Category	varchar	75	Used for quick identification of a group of files and folders that were examined together: Example 'Quick Categories': • Windows 11 Professional (64bit) – French; • Windows 11 Professional (64bit) –

			Russian; • Windows 10 Enterprise (64bit) - Chinese Simplified; • FreeBSD 4.6 (32bit); • Web Application; • Mac App Store; • Windows App Store;
File_Notes	varchar	255	Internal Use Only: Analysis notes we mentioned internally during our analysis.
Major	varchar	75	Main grouping label of a file or folder. Some Example of 'Majors': • Operating Systems; • Applications; • Etc.
Minor	varchar	75	Secondary or subgrouping from the file or folder's 'Major'. Some Example of 'Minors': • Installation; • Software Updates & Fixes; • Decompressed and/or Extracted; • Etc.
Operating_System	varchar	75	Name of the affiliated operating system, if applicable. Some examples: Windows Linux macOS

Manufacturer	varchar	150	Name of the manufacturer, if known or applicable.
Version	varchar	50	Version name, if known or applicable.
Is_Deleted	char	7	If the file or folder was found to be deleted by the host operating system, if applicable. Will be marked with either 'True' or 'False'.
Website	varchar	75	Website source, if applicable.
Geographic_Location	varchar	50	Geographic location of the manufacturer's Headquarters. If not known or not applicable then manufacturer's intended audience. Will be marked with either: • North America • South America • Europe • Asia • Asia/Pacific • Middle East • North America/Europe/Asia • Worldwide • Undetermined Potentially other regions or regional groups will be added in the future.
Extraneous	char	7	(Work in Progress) Intended to identify ignorable files that are unique to one particular situation. For example system log files, registry files, etc.

			Will be marked with either 'Yes' or 'Unknown'.
Log	char	7	(Work in Progress) Intended to identify potential log files. Will be marked with either 'Yes' or 'Unknown'.
Bad_Extension	varchar	255	File found to have a suspected bad file extension in comparison to the file's signature/header information. Will be marked with either 'True', 'False' or 'Unknown'.
actual_file	varchar	255	True if an actual file. False if derived or generated from an actual file (e.g. data broken out from compound files, EXIF data from graphic images, file metadata, and so on. Will be marked with either 'True', 'False' or 'Unknown'.
file_class	varchar	255	Class of file such as Regular File, Symbolic Link, etc, if known.
folder	varchar	255	Identified if a Folder. Will be marked with either 'True' or 'False'.
category	varchar	255	Specific type of file, if available, such as EXE, Text, Unicode, 7 bit text, etc.
compressed	varchar	255	Found as a compressed file, if applicable.
compressed_file_size	bigint	15	The compressed size of the file in bytes (compressed files only).
Compression_Method	varchar	255	Compressed file's compression

			method (Zip files only) such as Deflated, Stored, etc.
Extract_Version	varchar	255	Compressed file's extraction version (Zip files only), if identifiable.
permissions	varchar	255	Primarily UNIX like read/write/execute permissions, if identifiable.
UID	int	255	Primarily UNIX like User ID, if identifiable.
Group_Name_UNIX	varchar	255	UNIX like Group Name, if identifiable.
GID	int	255	UNIX like Group ID, if identifiable.
Username	varchar	255	Username of the file (Unix file systems only), if identifiable.
Container	varchar	255	Does the file, disk, partition, etc, have any children. Will be marked with either 'True' or 'False'.
Encrypted	varchar	255	Found to have Encryption file like qualities.
Deleted_Date	datetime	0	The Date deleted (Unix file systems only).
checksum	varchar	255	Checksum computed on compressed files (Zip files only).
Hash_Search_Engine_Record_Date	date	0	The date we added the file or folder's data into our databases.
Full_Path_Forensic_View	text	0	(Work in Progress) The full path to the file or folder from an operating system's perspective. Specialized computer forensic software makes this viewable.

5. In Summary

If, for whatever reason, you run into any issues or problems understanding the database table structures, data or field types then please feel free to reach out to us via our website HashSets.com. We will make every effort to provide you with a reasonable amount of additional information to help you better understand the database tables.