# Data Formats
# HashSets.com Databases

**Please Note: Within this document we have removed some older and insignificant table fields to help you in simplifying table importation and subsequent table data analysis. Please note the changes further below and implement into your own new or preexisting database tables.**

Updated: 12 October 2023

## 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 698+ operating system versions we have installed and forensically analyzed since 2003. The process involves installing the operating systems onto sterile hard drives and/or virtual disks, performing analysis using commercially available computer forensic software to discover all files, file attributes and meta-data and most importantly the MD5, SHA1, SHA-256 and SHA-512 hash values.

All of the aforementioned findings are then imported into individual operating system tables which subsequently form our HashSets.com database and online search engine. We concurrently offer to our Gold and Platinum subscription members the database tables exported into large tab delimited text files which may be reimported into their own database or third-party software products. Each table also provides a header row for simple field name explanation and identification (see further below for details).

## 2. Database Overview of the 'Primary Database Export' Table

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

The '**Primary Database Export**' table contains file information that never changes including the individual file's corresponding MD5, SHA1, SHA-256 and SHA-512 hash values, the file's first 32 bytes in hexadecimal, the first 128 ASCII characters (similar to using Unix or Linux 'strings' command), logical byte file size, and so on.

The '**Primary Database Export**' table contains hash values of safe, non-threatening, known-good computer files only. There are also NO duplicate hash values in this table. This is considered the master table that would be associated with the remaining fourteen tables discussed further below.

## ‘Primary Database Export’ Table Schema

| Field Name | Type | Length | Description |
|---|---|---|---|
| **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.<br><br>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.<br><br>Will be marked with either 'Yes' or 'Unknown'. |
| **Operating_System_File** | char | 7 | Hash values found within an Operating System.<br><br>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.<br><br>Will be marked with either 'Yes' or 'Unknown'. |
| **MS_Windows_OS** | char | 7 | Hash values found within Microsoft Windows operating system installations and updates.<br><br>Will be marked with either 'Yes' or 'Unknown'. |
| **Linux_OS** | char | 7 | Hash values found within Linux operating system installations and updates.<br><br>Will be marked with either 'Yes' or 'Unknown'. |
| **BSD_OS** | char | 7 | Hash values found within BSD (UNIX like operating system) installations and updates.<br><br>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.<br><br>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.<br><br>Will be marked with either 'Yes' or 'Unknown'. |
| **Gold_Disks** | char | 7 | Hash values found within Oracle's (formerly Sun) Solaris operating system installations and updates.<br><br>Will be marked with either 'Yes' or 'Unknown'. |
| **Installation_Discs** | Char | 7 | Hash values found within operating system installation media (CD, DVD and .ISO) files.<br><br>Will be marked with either 'Yes' or 'Unknown'. |

## 3. All Other Remaining Tables Described

The remaining fourteen tables are divided into operating systems (MS Windows, Linux, BSD, macOS and Solaris) and non-operating system groups (MS Windows App Store, Mac App Store, Gold Disks, Installation Media, third-party manufacturer's Software & Hardware Applications, etc).

Below is a description of each of the fourteen remaining tables. It should be mentioned briefly that over the course of many years (20+ years) the size of MS Windows and macOS operating system tables grew too large to contain within one table each. Therefore, in July 2021, we decided to divide those tables into smaller tables for convenience and ease of importation. The MS Windows tables were ultimately divided into three smaller tables (Windows North America, Windows Europe, Windows Asia and Windows Middle-East). For macOS we divided the previous single table into two tables ('macOS versions 8 thru 10' and 'macOS versions 11 and above').

- **Windows_North_America_Database_Export** – This database table contains various Microsoft Windows operating system versions for the United States, Canada (French Canadian), Mexico (Spanish Mexican) and the individual file or folder names, file name extensions, 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.

- **Windows_Europe_Database_Export** – This database table contains various Microsoft Windows operating system versions for Europe (Western and Eastern) and the individual file or folder names, file name extensions, 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.

- **Windows_Asia_Database_Export** – This database table contains various Microsoft Windows operating system versions for Asia (Japan, China and so on) and the individual file or folder names, file name extensions, 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.

- **Windows_Middle_East_Database_Export** – This database table contains various Microsoft Windows operating system versions for the Middle East (Hebrew, etc) and the individual file or folder names, file name extensions, 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.

- **Linux_Database_Export** – This database table contains various Linux operating system distributions and the individual file or folder names, file name extensions, 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.

- **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, 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.

- **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, 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.

- **BSD_Database_Export** – This database table contains various BSD (UNIX like) operating system distributions and the individual file or folder names, file name extensions, 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_Database_Export** – This database table contains various Solaris operating system versions and the individual file or folder names, file name extensions, 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.

- **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 particular software, program, third-party utility, hardware device, etc.

## 4. 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".

Note: As mentioned at the beginning of this document the below table schema no longer includes older fields labeled 'Record_Date', 'Key_Field', 'From_Recycle_Bin', 'From_Free_Space', 'Moved_From_Location', 'Graphic' and 'Inode_Number'.

Note: Additionally, we have lengthened the below fields labeled 'Extraneous' and 'Log' from 3 to 7.

| 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.<br><br>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.<br><br>Uses the following Date and Time format:<br><br>DD/MM/YYYY Hour:Minute:Seconds<br><br>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.<br><br>Uses the following Date and Time format:<br><br>DD/MM/YYYY Hour:Minute:Seconds<br><br>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:<br><br>Example 'Quick Categories':<br><br>• Windows 11 Professional (64bit) – French; |

| | | | |
|---|---|---|---|
| | | | • Windows 11 Professional (64bit) – Russian;<br>• Windows 10 Enterprise (64bit) - Chinese Simplified;<br>• FreeBSD 4.6 (32bit);<br>• Web Application;<br>• Mac App Store;<br>• 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.<br><br>Some Example of 'Majors':<br>• Operating Systems;<br>• Applications;<br>• Etc. |
| **Minor** | varchar | 75 | Secondary or subgrouping from the file or folder's 'Major'.<br><br>Some Example of 'Minors':<br>• Installation;<br>• Software Updates & Fixes;<br>• Decompressed and/or Extracted;<br>• Etc. |
| **Operating_System** | varchar | 75 | Name of the affiliated operating system, if applicable.<br><br>Some examples: |

| | | | |
|---|---|---|---|
| | | | Windows<br>Linux<br>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.<br><br>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.<br><br>Will be marked with either:<br><br>• North America<br>• South America<br>• Europe<br>• Asia<br>• Asia/Pacific<br>• Middle East<br>• North America/Europe/Asia<br>• Worldwide<br>• Undetermined<br><br>Potentially other regions or regional groups will be added in the future. |
| **Extraneous** | char | 7 | (Work in Progress)<br>Intended to identify ignorable files that are unique to one particular situation or instance |

| | | | |
|---|---|---|---|
| | | | For example system log files, registry files, etc.<br><br>Will be marked with either 'Yes' or 'Unknown'. |
| **Log** | char | 7 | (Work in Progress)<br>Intended to identify potential log files.<br><br>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.<br><br>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.<br><br>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.<br><br>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.<br><br>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. |

## 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.