DuraCloud Retrieval Tool

Introduction

The Retrieval Tool is a utility which is used to transfer (or "retrieve") digital content from DuraCloud to your local file system. It uses the command line (also called the terminal) on your local system to access your DuraCloud account and transfer the content you specify to a local file location of your choice. The content you retrieve also remains in DuraCloud; no content is deleted.

Familiarity with the command line will help you in using the retrieval tool:

- Windows command line tutorial
- Introduction to Mac OS X terminal

System Requirements

The Retrieval Tool has the same OS and Java requirements as the SyncTool. The system requirements for operating the SyncTool are described here. The Retrieval Tool also requires that there be sufficient disk space to retrieve the required content set from DuraCloud.

Download

Download the retrieval tool from the Downloads page.

Quick Start

For the impatient: Here are the commands you would use to complete the following common tasks. See the sections below for more details about how the Retrieval Tool operates and more information about available options.

You will need to replace the sections in { } !!

Download all files in a space

```
java -jar retrievaltool-{version}-driver.jar -h {your-duracloud-subdomain}.duracloud.org -u {your-username} -p
{your-password} -s {name-of-the-space-to-download} -c {name-of-local-directory-to-place-content}
```

Download a single file

First: Create a file, named content-list.txt, and place it next to the Retrieval Tool jar file. Open this file in a text editor and add to the first line the full content ID of the one file you want to download. You can find the content ID by logging into the DuraCloud UI and selecting the space. Each of the files that are listed as being in the space will be displayed with their full content ID. If your file has been chunked, remove the chunk extension (e.g. ".dura-manifest" or ".dura-chunk-0001") to get the content ID for use with the Retrieval Tool.

```
java -jar retrievaltool-{version}-driver.jar -h {your-duracloud-subdomain}.duracloud.org -u {your-username} -p
{your-password} -s {name-of-the-space-where-the-file-is-stored} -c {name-of-local-directory-to-place-content} --
list-file content-list.txt
```

Download a subset of files in a space

This is a two-part action.

First: Retrieve the list of all content items in the space:

```
java -jar retrievaltool-{version}-driver.jar -h {your-duracloud-subdomain}.duracloud.org -u {your-username} -p
{your-password} -s {name-of-the-space-to-list} -c {name-of-local-directory-to-place-list} --list-only
```

Second: Copy the list file created in the first step and paste it next to the retrieval tool, then open it and delete all content IDs in the list that you DO NOT want to download. When you are done you will have a list that contains only the files that you do want downloaded. Then start the download:

```
java -jar retrievaltool-{version}-driver.jar -h {your-duracloud-subdomain}.duracloud.org -u {your-username} -p
{your-password} -s {name-of-the-space-where-files-are-stored} -c {name-of-local-directory-to-place-content} --
list-file {name-of-the-list-file}
```

How the Retrieval Tool Works

- When the Retrieval Tool starts up, it connects to DuraCloud using the connection parameters you provide and gets a list of content items in the spaces you indicate. It will then proceed to download the files from those spaces, each into a local directory named for the space, which is placed within the content directory.
- For each content item, the Retrieval Tool checks to see if there is already a local file with the same name. If so, the checksums of the two files are compared to determine if the local file is the same as the file in DuraCloud. If they match, nothing is done, and the Retrieval Tool moves on to the next file. If they do not match, the file from DuraCloud is retrieved.
- By default, when a local file exists and differs from the DuraCloud copy, the local file is renamed prior to the DuraCloud file being retrieved. If you would prefer that the local file simply be overwritten, you will need to include the overwrite command-line flag when starting the Retrieval Tool.
- As each content file is downloaded, a checksum comparison is made to ensure that the downloaded file matches the file in DuraCloud. If the checksums do not match, the file is downloaded again. This re-download will occur up to 5 times. If the checksums still do not match after the fifth attempt, a failure is indicated in the output file.
- As each file download completes, a new line is added to the retrieval tool output file in the work directory, indicating whether the download was successful or not. Files which did not change are not included in the output file.
- As the Retrieval Tool runs, it will print its status approximately every 10 minutes to indicate how many files have been checked and downloaded.
- Once all files are retrieved, the Retrieval Tool will print its final status to the command line and exit.
- As files are updated in DuraCloud, you can re-run the Retrieval Tool using the same content directory, and only the files which have been added or updated since the last run of the tool will be downloaded.
- A file containing the list of content files within a space can be created using the "list-only" option (-l) instead of retrieving the actual content files themselves. The format of this text file is one content file name per line. This can be useful for many things.
- Specific content files can be retrieved from a space using the "list-file" option (-f) instead of retrieving all content files from a space. This can be useful by saving lots of time and bandwidth usage. One way to do this would be to first run a retrieval-tool command to create a file containing all content file names in a space using the "list-only" option. Then editing the text file containing the list of content names so it only contains a list of the desired content names and then use this file with the "list-file" option.

Operational notes

- Content Directory - the directory to which files will be downloaded. A new directory within the content directory will be created for each space.
- Work Directory - the work directory contains both logs, which give granular information about the process, and output files. An output file is created for each run of the Retrieval Tool which stores a listing of the files that were downloaded.

Prerequisites

As of DuraCloud version 7.0.0, the Retrieval Tool requires Java 11 to run. The latest version of Java 11 is available from AdoptOpenJDK.

- You must have Java version 11 or above installed on your local system. If Java is not installed, or if a previous version is installed, you will need to download and install Java 11. To determine if the correct version of Java is installed, open a terminal or command prompt and enter

```
java -version
```

  The version displayed should be 11.0.0 or above. If running this command generates an error, Java is likely not installed.
- You must have downloaded the Retrieval Tool. It is available as a link near the top of this page.

Using the Retrieval Tool

- To run the Retrieval Tool, open a terminal or command prompt and navigate to the directory where the Retrieval Tool jar file is located
- To display the help for the Retrieval Tool, run

```
java -jar retrievaltool-{version}-driver.jar
```

- When running the Retrieval Tool, you will need to use these options:

Short Option	Long Option	Argument Expected	Required	Description	Default Value (if optional)
-h	--host	Yes	Yes	The host address of the DuraCloud DuraStore application	
-r	--port	Yes	No	The port of the DuraCloud DuraStore application	443
-u	--username	Yes	Yes	The username necessary to perform writes to DuraStore	

-p	--password	Yes	No	The password necessary to perform writes to DuraStore. If not specified the retrieval tool will first check to see if an environment variable named "DURACLOUD_PASSWORD" exists, if it does exist the retrieval tool will use its value as the password, otherwise you will be prompted to enter the password.	Not set
-i	--store-id	Yes	No	The Store ID for the DuraCloud storage provider	The default store is used
-s	--spaces	Yes	No	The space or spaces from which content will be retrieved. Either this option or -a must be included	
-a	--all-spaces	No	No	Indicates that all spaces should be retrieved; if this option is included the -s option is ignored	Not set
-c	--content-dir	Yes	Yes	Retrieved content is stored in this local directory	
-w	--work-dir	Yes	No	Logs and output files will be stored in the work directory. If not specified, this value will default to a directory named duracloud-retrieval-work in the user's home directory.	duracloud-retrieval-work
-o	--overwrite	No	No	Indicates that existing local files which differ from files in DuraCloud under the same path and name sould be overwritten rather than copied	Not set
-t	--threads	Yes	No	The number of threads in the pool used to manage file transfers	3
-d	--disable-timestamps	No	No	Indicates that timestamp information found as content item properties in DuraCloud should not be applied to local files as they are retrieved.	Not set
-l	--list-only	No	No	Indicates that the retrieval tool should create a file listing the contents of the specified space rather than downloading the actual content files. The list file will be placed in the specified content directory. One list file will be created for each specified space.	Not set
-f	--list-file	Yes	No	Retrieve specific contents using content IDs in the specified file. The specified file should contain one content ID per line. This option can only operate on one space at a time.	Not set

Examples of running the retrieval tool:

- Retrieve all the files stored within the 2 specified spaces and place them in the specified local content directory under sub-directories matching the specified 2 space names.

```
java -jar retrievaltool-{version}-driver.jar -c content -h test.duracloud.org -u myname -p mypassword -s
space1 space2 -o
```

- Retrieve all the files stored within all spaces and place them in the specified local content directory under sub-directories matching the space names.

```
java -jar retrievaltool-{version}-driver.jar -c content -h test.duracloud.org -u myname -p mypassword -a
```

- Create a file containing the list of content IDs for the specified spaces using hidden password option (-p command line option not specified, will be prompted for password). This example would not actually retrieve the content files, rather it creates a list of content files in the specified space. Each specified space will have its own content list file created in the specified local content directory. The naming convention of each list file created will be: "<space_id>-content-listing-<storage_provider>.txt"

```
java -jar retrievaltool-{version}-driver.jar -h <host> -u <user> -c <content_dir> -s <list of space IDs
separated by a space> -l
```

- Retrieve only the specified contents by using the list-file option (-f). The -f option can only operate on one space. This command would result in h aving all the content files listed in the specified file of content IDs placed in the specified local content directory.

```
java -jar retrievaltool-{version}-driver.jar -h <host> -u <user> -c <content_dir> -f
<path_to_file_of_specified_content_IDs> -s <single space ID>
```