Prudential – NICE Recording Archive

Author: Clifford Grimm

Date: 7/22/2013 5:26 PM

Revision: 0.5

DISCLAIMER: This document alone is not a formal contract between HDS and Prudential, but is intended to describe the proposal for the deliverable.

Table of Contents

1. [Revision History 2](#_Toc362270700)
2. [Introduction 3](#_Toc362270701)
3. [Delivery Details 4](#_Toc362270702)
4. [Custom Metadata Extraction Details 8](#_Toc362270703)
5. [Known Problem Areas 10](#_Toc362270704)
6. [Risk/Concerns 10](#_Toc362270705)
7. [Future Considerations 11](#_Toc362270706)

# Revision History

|  |  |  |  |
| --- | --- | --- | --- |
| **Date** | **Author** | **Revision** | **Description** |
| 6/3/2013 | Clifford Grimm | 0.1 | Initial Revision |
| 6/4/2013 | Clifford Grimm | 0.2 | Grammar/Typo fixes based on feedback. |
| 6/5/2013 | Clifford Grimm | 0.3 | Updates based on call with Prudential. Time zone now known. |
| 7/9/2013 | Clifford Grimm | 0.4 | Updates based on latest info and design direction. Added current known file counts and risk/speed assessments. Still need final audio file numbers from HDD7. |
| 7/22/2013 | Clifford Grimm | 0.5 | Added more information about problem files on HDD1 and HDD2. |

# Introduction

Prudential has approximately 16TB of NICE call recordings that were exported from the application and stored on USB drives. There are concerns about the life of the USB drives and need to move them to alternative storage. In addition, the current mechanism to locate call recordings is cumbersome and relies on manual searching through the call recording names to locate the appropriate call recording.

The original storage of the call recordings to USB was accomplished by a 3rd party vendor, Sykes, that extracted the recordings from NICE, formulated a folder and file name structure to identify the call, then stored them on the USB drives.

To accomplish both goals of safe storage and search, HCP will be used as a repository for these call recordings. HCP was purposely designed to host unstructured data like call recordings. The HCP Search Console, MQE and Annotation capabilities in HCP 6.0 will be used to enable Prudential to locate content based on a range of criteria like the agent and date characteristics.

This delivery will focus on the mechanism to ingest the call recordings into HCP with custom metadata to reach the storage and search goals.

# Delivery Details

The delivery will be a joint effort between Prudential and HDS. As an overview, Prudential will be responsible for transferring the call recordings to HCP with guidance from HDS. HDS will be responsible for adding metadata to the call recordings to allow for robust search and providing basic training for use of the search capabilities.

The following picture is an overview of the approach.



HCP-DM via REST

COMET Tool

Scan files (CIFS/NFS)

Add Custom Metadata

(REST)



* Perform Audit
* Search and Remove Driver Files



Tool Log DB

Tool Log File

Web Access



Search for Calls

Web Access

Virtual or Physical Machine

Figure 1: Delivery Overview

With guidance from HDS, Prudential will move the content from the USB Drives to an HCP Namespace using the HCP-DM tool. HCP-DM is a multi-threaded application that can migrate content to HCP in the most efficient manner. Prudential will be responsible to ensure that the content on the USB drives have been correctly written to HCP using any level of content verification they deem appropriate.

The migration may take a week or more to accomplish as the expected bandwidth of the whole process will likely be limited by the USB drive. In default mode, HCP-DM executes 20 threads per HCP node to achieve the most efficient HCP ingest; all the HCP-DM threads will be accessing the USB drive at the same time causing excessive disk thrashing. It might be beneficial to throttle back HCP-DM to only have a few simultaneous threads.

During initial start of migration, the file/folder layout and a sample of metadata files were examined to establish a base understanding of the actual content. Initial conversations with Prudential indicated that the information of Agent ID, Logger, Start Time, and End Time is sufficient for search criteria.

HDS will develop specialized modules for Prudential that will understand how to build custom metadata based on the content uploaded to HCP from the USB drive. These specialized modules will be used with a general purpose tool called COMET (Custom Metadata Extraction Tool). This tool has the logic for iterating over content, calling the extraction module, and writing the generated custom metadata to HCP.

Since the content will have been migrated to HCP by HCP-DM, the COMET tool will access the content in the HCP namespace using either CIFS/NFS. It will walk the content using these protocols and add the custom metadata via the HTTP REST interface. After analyzing one of the two USB drives content, the following is the expected custom metadata that will be added to each Aud/WAV file stored on HCP.

<CallRecording type=”Aud|wav”>

<Logger>#######</Logger>

<Channel>######</Channel>

<Server>####</Server>

<Session>#####</Session>

<CLSCallID>####</CLSCallID>

<AgentID>####</AgentID>

<DateTime>

<Start>yyyy-MM-ddTHH:mm:ssZ</Start>

<End>yyyy-MM-ddTHH:mm:ssZ</End>

<DateTime>

</CallRecording>

Figure 2: Basic Custom Metadata Layout

Special Notes about Custom Metadata:

1. There is no Agent ID information on the earlier recordings. If Agent ID is required search criteria, a mechanism must be identified and developed to provide this information during ingest. An alternative would be to have a second pass at the content and add the Agent ID based on some external source of mapping and the fields that are available (e.g. Server/Session). Neither of these approaches are part of the scope of this project at this time.
2. The start and end date-time fields will be stored in UTC time. With the HCP MQE search engine, it will be possible to specify any time zone in the search criteria and the appropriate interpretation will be performed automatically by HCP.
3. The dates obtained from data on USB drives are in Eastern Standard Time (EST) zone.

Before performing the full comprehensive run to generate the Custom Metadata, HDS will perform an isolated sample run of the COMET tool to validate proper processing. The samples will be chosen as to represent each file/folder layout as written to HCP from the USB drive.

The sample run will consist of the following steps:

1. Define the HCP Content Class to represent the Custom Metadata described above. The HCP Content Class is the mechanism to define what of the Custom Metadata should be indexed, the data type for that content, and a user consumable name for the fields.
2. Select samples of content on HCP based on folder. Ensure each sample selected is a representative of each of the many file layouts.
3. Run COMET tool against the samples and wait for the indexing of the new custom metadata.
4. Spot check a 2-5 pieces of content from each sample verifying that the Custom Metadata is valid.
5. Perform simple searches via the HCP Search Console to ensure content is properly indexed.

After validation of the sample run, HDS will initiate a full COMET tool run against the NICE recordings. This can either be performed against the entire HCP Namespace, or can perform simultaneous tool execution against separate “HDD” folders. If performing simultaneous tool execution, care will need to be taken to ensure there are multiple tool environments to avoid configuration or log file collisions. In addition, each tool instance is recommended to execute against a separate CIFS/NFS mount point to the HCP Namespace using separate HCP nodes; this configuration will help with throughput and concurrency.

Monitoring of the tool processing will primarily be performed by Prudential personnel with any required assistance remotely by HDS. Any requirement for a secondary on-site visit can be discussed between the companies.

Final analysis will be performed by HDS and Prudential in the following manner:

1. Search for all Aud/WAV files that do not have Custom Metadata. Confirm reason is due to no information available for NICE recording files. Resolve any issues where information is available but not generated.
2. Perform a spot check of up to 10 Aud/WAV files for each layout that exists from the USB drives to ensure the proper mapping was achieved.
3. Once indexing has been complete, perform searches against generated Custom Metadata to ensure proper indexing.
   1. Perform a date based search on date/time fields.
   2. Perform search based on other field like Logger.

For validation purposes, HDS will provide all log files produced by the COMET tool and provide an HCP generated query result providing evidence of the expected Aud/WAV files have Custom Metadata attached.

A brief training session will be provided either via WebEx or on-site for the core consumer of the HCP search capabilities and can occur any time after some content has been indexed. The expected duration of the training is likely to be about an hour. The training will cover the basics of performing searches using the HCP Search Console against the actual NICE recording content. If necessary, the training may also consist of developer focused training for utilizing the API to perform content searches.

# Custom Metadata Extraction Details

For scoping this effort, an analysis of the file layout of the USB drives was performed and a sampling of the “downloader” files was examined. Upon analysis, there was found to be a wide range of data layouts with only mild consistency. In general, there were the following observations:

1. The folder structure that contains the Aud/WAV files varied too widely to rely on for any information. Luckily, the folder structure does not contain core information required for generating the Custom Metadata.
2. There are two core mechanisms identified for generating the Custom Metadata for indexing and search capabilities:
   1. Aud/WAV file name contains all information. The audio file layout is for the following form:

SC\_*<logger>*\_*<start-date>*\_*<start-time>*\_*<end-date>*\_*<end-time>* \_*<channel>*\_*<CLSCallID>*\_*<AgentID>*.Aud [ File Count: 11,625,741 ]

***NOTE***: The listing for HDD7 was incomplete due to drive failure. This above number will need to be updated when the full listing of HDD7 is provided.

**Failure Risk:** **Low**

**Processing Speed: High**

* 1. There is a “Prudownloader” mapping file concept that consists of a text file containing a mapping from Aud/WAV file name to its information that will be used to generate the Custom Metadata. The layout of this file is not straight forward to parse, so great care will have to be taken.  
       
     In addition, there are two ways that the folder and filename for the mapping file can be determined in relation to the location of the Aud/WAV files.  
     1. In the same folder as the Aud/WAV file, there is a companion “prudownloader\_audfiles.txt”.

Audio file names:

* sykespru\_*<logger>\**.wav [ File count: 79,905 ]
* sykespru\_WAV\*.wav [ File count: 43 ]
* Prudownloader\_Pru\*.wav,  
  prudownloader\_Pru\*.wav [ File count: 1,201,524 ]
* Prudownloader\_WAV\*.wav [ File count: 50 ]
* Prudownloader\_Pru\*.aud [ File count: 150 ]

**Failure Risk:** **Low**

**Processing Speed: Medium/Low**

* + 1. Somewhere nearby in the folder structure of the “HDD” top-level folder, there is a companion “Prudownloader” file. Finding the folder and the precise mapping file will be challenging due to the following:
       1. The folder structure and relation to the Aud/WAV file folder is not consistent. Hopefully, there is a small enough variation to be possible for hard-coding into the tool.
       2. The name of the mapping file is not consistent except that it starts with “Prudownloader”, has “audfile” somewhere, and is a (.txt) file. It is currently unclear if an efficient algorithm can be devised to collect the Custom Metadata without searching through all files.

Audio file names:

* *<logger>*\_*<sdate>*\_*<stime>*\_\*.wav [ File count: 698,680 ]
* *<logger>*-*<date-mmddyy>*\_\*.wav [ File count: 1,620 ]

**Failure Risk:** **Medium/High**

**Processing Speed: Low**

Special Notes/Questions:

1. There are a number of files for which there does not appear to be any metadata mapping capabilities. See section *Known Problem Areas* for the details of the findings.
2. Any Custom Metadata values that cannot be devised via these mechanisms will be left blank in the attached Custom Metadata.

# Known Problem Areas

During detail analysis of the file listings from the Hard Disk Drives (HDD), a number of problem areas were identified. This will impact the ability to associate metadata in HCP to audio files for enabling search and discovery.

1. On HDD1
   1. There are ***151,454 files*** in 14 tape folders that does not have any mapping file. The only useful information that can be retrieved from file name is the logger, start/end time/date, and channel number.   
      The following is a list of the root of the folders for which this applies:
      1. HDD1\PC1\Tape 10
      2. Rooted at HDD1\PC2: TAPE1, TAPE2, TAPE 6  through TAPE16
   2. Unexpected file name for the following file as it does not match the file naming pattern for other files in that folder:
      1. HDD1\PC2\TAPE10\TAPE10-day2\Prudownloader\_Prudential call WAV day 2\_377A.wav
2. On HDD2, there is a folder that contains ***a 262 mapping files*** that is not obvious as to what they go to.  All other folders with audio files on this HDD has some kind of mapping file either in the folder with the audio file, or has a corresponding folder with mapping files.   
   The path to the folder is: HDD # 2\Prudential downloader aud file
3. On HDD3
   1. There are ***245,637 files*** in 19 tape folders that do not have any mapping file. The only useful information that can be retrieved from file name is the logger, start/end time/date, and channel number.   
      The following is a list of the folders that this appies applies:
      1. Rooted at HDD # 3: TAPE1, Tape 17 through Tape 34
4. On HDD6
   1. There appears to be a misnamed file:
      1. HDD # 6\DAY27\_APR24\_2012\AM\JAN12-17\_2007\1-12-07\11220201-011207\_D6AE (986).txt
   2. There are ***1,620 files*** found in the following folders that do not have any mapping file:
      1. HDD # 6\DAY27\_APR24\_2012\AM\JAN12-17\_2007\1-12-07
      2. HDD # 6\DAY27\_APR24\_2012\AM\JAN12-17\_2007\1-13-07

# Risk/Concerns

* New content layout is found on additional USB drives to be imported into HCP, thus requiring more development work for modules to handle this additional layout. Will need a full file listing of all USB drives that will be imported into HCP.
* Due to the inconsistent layout of the “downloader” files, it may be very difficult to provide an efficient algorithm to locate the appropriate “downloader” file, and extract the information for a specific Aud/WAV file. This may result in much higher compute/disk resources and thus may cause longer completion of the overall processing.
* The call recording metadata discovered and stored in the Custom Metadata is not sufficient for their search desires. Will need to ensure Prudential understands what information is available early and if insufficient work with Sykes to determine an appropriate way to obtain it. Any additional processing required beyond what is outlined above will require additional development work.

# Future Considerations

This effort does not include any HCP Retention Period settings on individual objects. All objects ingested into HCP will take the default configuration on the HCP Namespace; that is defaulted to “Deletion Allowed”. If it is desired to put retention on the individual objects based on when the call occurred as stored in the Custom Metadata, this will require another engagement with HDS or for Prudential to write a tool to traverse through the content and set an appropriate retention period.

For those audio files that do not have Agent ID information available, it may be possible to engage a second time and develop a mechanism to add the Agent ID information to the existing Custom Metadata.