HP Integrated Archive Platform
User Repository Utilization Report
Technical Addendum

This technical addendum describes the User Repository Utilization Report feature, its use, and its limitations.

Part Number: UtilReport
IAP 2.0.5/2.1.2 edition: November, 2010

Legal and notice information

© Copyright 2010 Hewlett-Packard Development Company, L.P.

About the User Repository Utilization Report

The User Repository Utilization Report feature provides you with a snapshot of the current utilization of IAP user repositories. A collection job gathers the utilization data and places it in an internal database. You can then download a ZIP archive that contains CSV-formatted User Repository Utilization reports and a collection job summary file. You can also configure the feature to send the ZIP archive into a valid IAP repository for historical documentation.

User Repository Utilization Report data

A CSV-formatted User Repository Utilization report contains the following information for each repository within an IAP domain. Some fields are only specific to certain classes of repositories.

Fields reported for all repository classes

- **REPOSITORY_CLASS** – There are three classes of repositories: user repositories, quarantine repositories, and system repositories. Every IAP user has a primary repository that belongs to that user. Users can quarantine a query result set by creating a quarantine repository and routing the query results into it. Repositories that are not associated with users are referred to as system repositories.
- **REPOSITORY_ID** – The unique identifier for the repository.
- **NUM_DOCS** – The total number of documents routed to the repository.
- **UNCOMPRESSED_SIZE** – The total uncompressed size (in bytes) of all documents routed to the repository. Each document is internally stored in a compressed archive. The information in this field conveys the total inflated size of all documents. The uncompressed size disregards the fact that documents may be routed to multiple repositories (single-instanced).[1]
- **COMPRESSED_SIZE** – The total compressed size (in bytes) of all documents routed to the repository. Each document is internally stored in a compressed archive. This information conveys the total amount of disk space that all of the repository's documents are using. The compressed size disregards the fact that documents may be routed to multiple repositories (single-instanced).[1]

Fields reported for user and quarantine repositories

- **USERNAME** – For users that are created in the PCC Account Manager (known as local users), the administrator manually enters the username. For users that are imported into the IAP via Dynamic Account Synchronization (DAS), the username is populated with a value from the LDAP server (Active Directory/Domino).
- **EMAIL_ADDRESS** – The user's email address as entered into the Email Contact field in the Account Manager.
- **LAST_NAME** – The user's last name as entered into the Last Name field in the Account Manager.
- **FIRST_NAME** – The user's first name as entered into the First Name field in the Account Manager.
- **OU** – The organizational unit to which the user belongs. This field is retrieved from the LDAP distinguished name field. Because LDAP configurations vary widely across companies, this field may not represent the organizational hierarchy for all customers.

[1] Single instancing means that only one physical copy is stored, but the appearance of multiple logical copies is provided. For example, an email that is sent to user1 and user2 generates a single document (compressed archive) inside the IAP, but is "stored" in both users' repositories. Disregarding single instancing means that the one physical copy is counted for each reference.

Fields reported for quarantine and system repositories

REPOSITORY_NAME – Every repository in the IAP has a name.

- For quarantine repositories, the name is based on the saved query results name as given by the IAP user. The naming convention is `<saved_query_name>.<repositoryid>.quarantine.repository`.
- For system repositories, the name is human readable, such as the domain Catchall repository; for example, `domain1.catch all`.

Job summary file

The User Repository Utilization report also contains a collection job summary file.

This file is the last part of the downloaded ZIP archive. If there is a problem downloading the report, the job summary file is dropped. Therefore, the existence of the job summary file indicates that the report has been successfully downloaded. (However, you should always examine the report for correctness when the download is complete.)

Fields reported in the job summary file

- **Domain** – The IAP domain on which the report was generated.
- **Ignore Documents Stored After** – The report excludes documents stored after this date. Generally this field shows the time that the utilization collection job was started. However, you can set it to some time in the past to create backdated reports.
- **Start** – The time that the utilization data collection job started.
- **End** – The time that the utilization data collection job ended.
- **Results** – Either *complete* or *incomplete*. The status is complete if the collection job completed without any errors. Otherwise, it is incomplete. If the collection job encountered an error, such as a Smartcell (storage node) not responding, the report can still be downloaded. In that case, the completeness of the report is unknown.
- **SmartCell Groups Expected** – The number of Smartcells that were expected to respond to the utilization query.
- **SmartCell Groups Responded** – The number of Smartcells that responded to the utilization query. If this number is smaller than the number expected, the report is incomplete. However, if this number is the same as the number expected it does not mean that the results are complete. Other errors might have been encountered to make the results incomplete.

Job summary modes

You can run the job summary feature in two different modes – repository mode or folder mode. In repository mode, the report collects utilization summary data for all documents in all repositories in the domain. In folder mode, the report collects utilization summary data of only those documents in user repositories that have attached folder information.

To enable the feature:

1. Open the `Domain.jcml` configuration file on the kickstart server and identify the domain for the utilization summary report.
2. Add the attribute `UtilizationReportMode`.
3. Set the value of `UtilizationReportMode`:
   - To enable the feature in repository mode, set the value to `repository`.
     Example: `UtilizationReportMode=repository`
   - To enable the feature in folder mode, set the value to `folder`.
     Example: `UtilizationReportMode=folder`
4. After saving and closing the `Domain.jcml` file, run `regloader.pl -cv -clearallconfirm=<IAP name>` on the kickstart server to finalize the configuration.
5. Run `/opt/bin/restart` from the PCC server to restart the IAP and update all servers.

Using the User Repository Utilization Report

The User Repository Utilization Report is generated in two steps. The first step is to collect the utilization data for all repositories in a domain. The second step is to download the report. There is a Web page for each step: the collection page and the download page.

You can access the User Repository Utilization Report from the left menu of PCC Web Administration by selecting **Reporting** > **Utilization Report**.

Utilization Report Collection page

The first page in the Utilization Report view is the Utilization Report Collection page. This page displays the status of the latest collection jobs and contains the mechanisms for triggering new collection jobs. It consists of four areas: tabbed IAP domains, Latest Collection Job Status, Collect Utilization Data, and Recurring Collection Jobs.

Tabbed IAP domains

Select the IAP domain by clicking the associated tab. The selected tag is colored gray.

Latest Collection Job Status

The collection job queries all the Smartcells (storage nodes) in the specified domain, aggregates the results, and places them in a database. In general, this process can take up to six hours to complete. However, a large-scale volume of users, with a corresponding volume of emails, will lengthen the processing time for the report.

You can monitor the job status in this area of the page. A job can be in one of the following states:

- **Running** – The job is currently running.
- **Completed** – The job has completed and the data has been placed in the database.
- **Completed with errors** – The job has completed, but it encountered an error during processing. This happens when only a subset of the Smartcells responded to the utilization query.
- **Failed** – An unrecoverable error occurred and the job was aborted.

You can perform the following actions:

- Get a report of completed jobs by clicking **Download Report**. After clicking this button, you are taken to the Report Download page.
- Archive completed jobs into the IAP by clicking **Archive Report**. The report will be attached to an email and sent into an IAP repository. The archive feature allows you to keep an historical record of the utilization reports, which can be retrieved via a search in the IAP Web Interface. An IAP repository for the reports must be selected before the report can be archived. (See Configure Archiving Parameters).
- Refresh the status of the latest collection job by selecting **Reporting** > **Utilization Report** in the PCC left menu.

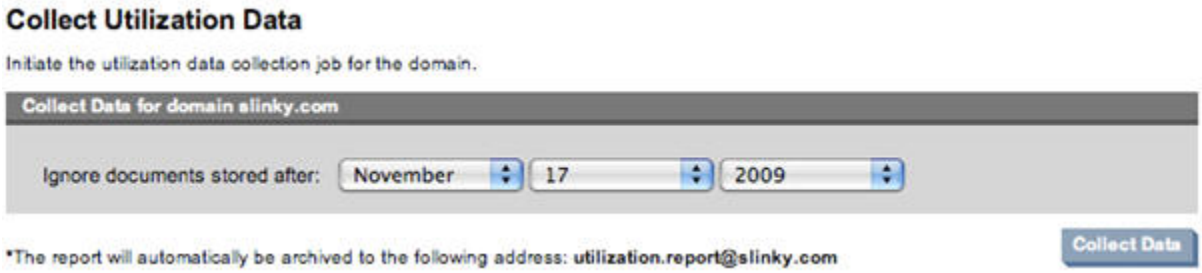
NOTE:

Be careful not to refresh the page using the Web browser's refresh option. You might unintentionally resubmit the last command (`collect data/archive report/schedule/delete/configure`). If this happens and the browser prompts you to continue resubmitting the form, select **Cancel**.

If a collection job encounters an error or fails, an SNMP trap is thrown. The traps, which are configured in the PCC SNMP Management page, can be sent to an external server or to an email address. The following traps have been added for utilization reports:

- **utilization_report_failure** – The collection job failed.
- **utilization_report_error** – The collection job completed with errors.
- **utilization_report_archiving_failure** – The utilization report failed to archive.

Collect Utilization Data



Use this area to run a single collection job.

1. Select the domain on which the collection job will be run in the tabbed IAP domains area at the top of the page.
2. If you want the report to be archived when the collection job is complete, configure the archiving parameters.
3. If necessary, configure the **Ignore documents stored after** field.

   The collection job ignores documents that are archived after the given date. By default, the job is set to ignore objects stored after midnight (11:59:59 p.m.) on the current day.

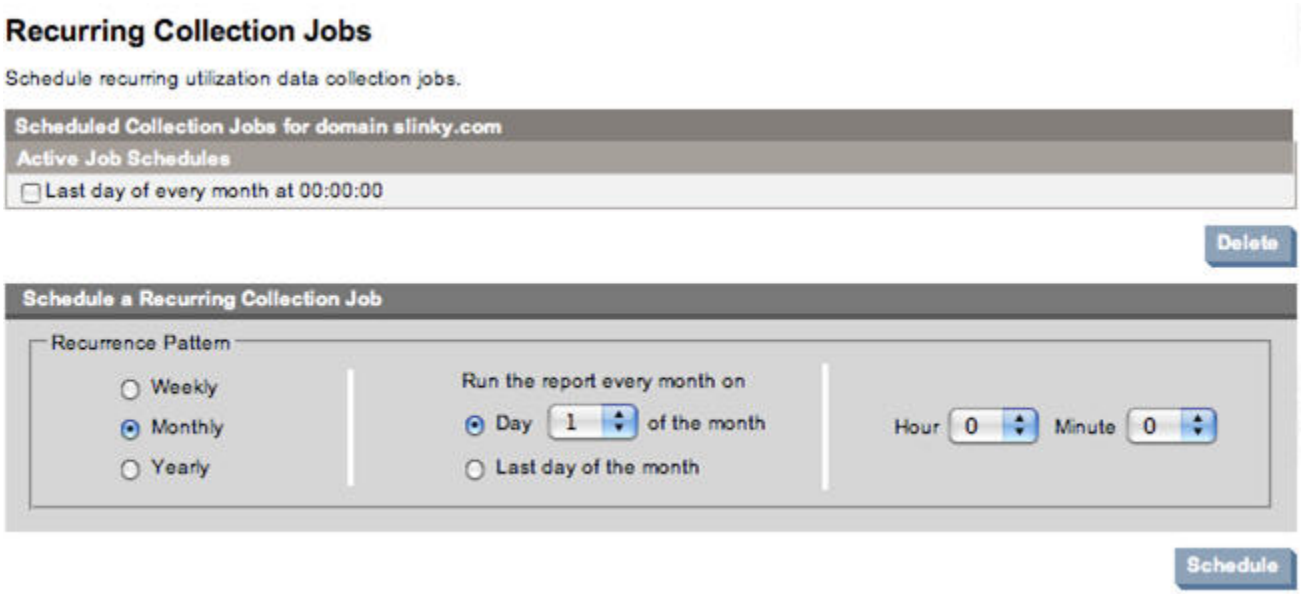
   You can set this field to an earlier date. This will generate a backdated report for all data currently stored in the domain that was ingested before the given date. Current document routing is applied.

   The backdated report feature should be used judiciously. The data collected in backdated reports can misrepresent the actual utilization on the selected date if the retention, reprocessing, or quarantine features have modified the document routing information. The retention feature can remove items from a repository, so that lower utilization data could be reported. The reprocessing feature routes email into a repository, so that higher utilization data could be reported. The quarantine feature creates and routes the document to an additional repository, which will be displayed in the report.
4. Start the collection job by clicking **Collect Data**.

   After the job has started, the Collect Data button becomes disabled.

   Only one collection job can run on a selected domain at a time. While the collection job is running, the status in the Latest Collection Job Status area of the page displays *running*.

Recurring Collection Jobs



You can schedule utilization collection jobs to run on a weekly, monthly, or yearly basis by following these steps:

1. Select the domain on which the collection job will be run in the tabbed IAP domains area at the top of the page.
2. If you want the report to be archived when a collection job is complete, configure the archiving parameters. Be sure to select the **Enable automatic archiving** check box.
3. Configure the recurrence pattern.
4. Click **Schedule**.

   The scheduled job will be listed in Active Job Schedules.

   (To remove a scheduled collection job, select the job in Active Job Schedules and click **Delete**.)

Configure Archiving Parameters

You can archive utilization reports for historical reference. The report is attached to an email and sent into an IAP repository. It can be retrieved through a search in the IAP Web Interface.

To configure the archiving feature:

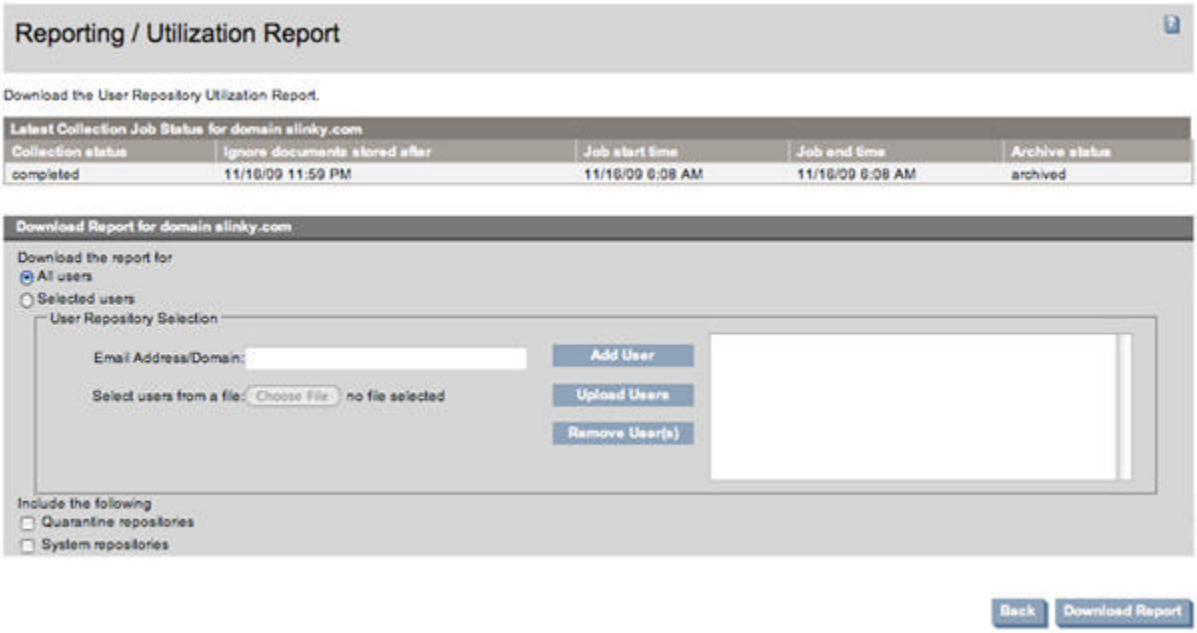
1. Ensure the relevant domain is selected in the tabbed IAP domains area at the top of the page.
2. In the text box, enter an email address that is associated with a valid IAP repository.

   You might want to create a repository in the domain especially for these reports.

3. If you want the utilization report to be archived automatically, select the **Enable automatic archiving** check box.
4. Click **Configure**.

Report Download page

You can download the User Repository Utilization Report for a selected collection job. The downloaded report is a ZIP archive containing CSV files and a report summary page. The data is chunked into CSV files with 65,000 lines (plus one header line) in order to maximize compatibility with Microsoft Excel.



To download a User Repository Utilization Report:

1. Click **Download Report** on the Utilization Report Collection page.
2. To download the report for all user repositories in the IAP domain, click **All users**.
3. To download the report for a select set of users, click **Selected users** and complete the User Repository Selection fields.
   - To select a single user, add the user's email address (for example, john.doe@domain.com) to the **Email Address/Domain** field and click **Add User**.
   - To select all users in an email domain, add the domain (for example, @domain.com) to the **Email Address/Domain** field and click **Add User**.

NOTE:

Downloading the report for all users in a selected email domain (such as @domain.com) requires the system to perform additional work to expand the list of users. Downloading for a large selection of users means that database queries are less efficient. Depending on the number of users to which the email domain expands, the download time can be greater than downloading the report for all users.

- Users can also be added by uploading a newline-separated file.[2] Click **Choose file**, select the file, and then click **Upload Users**.

The users are added to the selection box. There is a limit of 500 entries in this box. You can remove a user by selecting the email address and then clicking **Remove User(s)**.

4. Select the check boxes for **Quarantine repositories** and **System repositories** if you want to include them in the report.

   Quarantine repositories and system repositories can only be included if reports are in repository mode. (See Job summary modes.)

5. Click **Download Report**.

Limitations

The following limitations apply to User Repository Utilization Reports:

- Only one collection job can run at a time on a given IAP domain.
- Only one set of utilization data can be stored in the IAP database at a given time for each IAP domain.
- In general, the process of collecting utilization data can take up to six hours to complete. However, a very large volume of users, with a corresponding volume of emails, will lengthen the processing time for the report.

  During the time it takes to process the data, failures might occur on a subset of the Smartcells in the IAP domain. The job will continue to process even if some of the Smartcells do not respond, thereby creating a partial report. The report can still be downloaded, but it will not accurately report the utilization data for the domain. No mechanism exists to reissue the collection job on that subset of Smartcells. Instead, the job must be run again across the entire domain.
- This feature runs on the PCC server. If the PCC server is restarted, any collection jobs that are currently running will fail, and you will need to start a new job. The Smartcells, however, will continue to process the previous jobs to completion.

  To prevent the Smartcells from running out of memory, we strongly recommend that you wait a day before starting a new collection job, to allow the previous jobs to finish.
- The utilization report feature relies on several new fields in the content indexes. In order for documents to be scanned for a report, they must be indexed after this feature is installed. For existing customers, that means documents need to be reindexed. It is prohibitively expensive to reindex full Smartcells (it can take weeks), and the IAP is not fully functional during that time. Therefore, HP does NOT recommend enabling this feature at sites with large quantities of previously stored data.

[2] The newline-separated file should be a plain text file (use Notepad or vim) where every line contains either an email address (john.doe@domain.com) or email domain (@domain.com). The newline-separated file can have any extension.