Email Classification & Document Data Mining
Phases 1-3 Generic Specifications

Introduction

The Client/Counsel department(s) have a large number of email documents stored in and around the organization in many different locations, systems and formats. The goal is to organize the information from these email messages and attachments and to create an ongoing and intuitive system that monitors and classifies the content, stores it in a logical folder schema, reports back on it and presents information up in an easy to use way. The work is divided into three phases, with increasing numbers of files, users, and timeframes at each step.

PHASE 1: PROOF OF CONCEPT

Objectives

- Identify, evaluate and classify select emails and attachments that are housed in the existing email system or DMS. Current email message volume estimate is [client-specific] messages. After de-duplication, expected volume for processing is [client-specific] email messages with [client-specific] attachments for a total doc count of [client-specific] files.
- Provide Proof of Concept and working model for limited environment typically consisting of 1-5 custodians, a partial folder or classification taxonomy, and limited document volume.
- Provide automated indexing, tagging, keywords and other key document metadata to all Phase One documents.
- Provide easy-to-use document analysis & data visualization tool for easy, intuitive document population assessment, organization and disposition for limited users.
- Run offline, outside of normal business hours and processes in Valora's protected staging environment or in Client's data center.

Phase One typically operates at Valora's Bedford, MA facilities or behind Client's firewall (or in Client's data center). When occurring at Valora, Client's IS group collects the files outlined above and sends to Valora on hard media or via secure ftp. Valora receives the documents, logging all chain of custody, and processes them in Bedford through PowerHouse. If data is to remain at Client site, Client's IS group provides remote access to Valora to run PowerHouse locally at Client site.

Once PowerHouse processing is complete and tested (may take several iterations), Valora uploads the documents and all indexing work product to Client's DMS and/or to BlackCat, which provides a dashboard view and other data visualization elements securely over the web. During Phase One, BlackCat is hosted either at Valora's facility or Client's data center. When hosting the Phase One documents inside BlackCat, Valora typically also delivers a complete set of Phase One database load files to client for loading to internal systems.

Plan Steps & Rough Timing

1. Client collects and produces guidance and specifications about how emails & attachments should be classified. Specification Example: File by client matter number. Guidance Example: list of all active case matters and attorneys/staff assigned to each.
2. Valora custom-configure PowerHouse ("PH") to client specifications. (Often overlaps with Step 1)
3. Client supplies test set(s) of custodian data to Valora.
4. Valora tests configuration against test data set.
5. Valora configures BlackCat and populates with test data set & results, concurrent with Step 4.
6. Client begins testing and feedback.

Timing for Phase One includes Phase One Setup & Configuration, prior to actual file processing. Phase One duration is typically 1-2 months.

Cost

There are three main fee areas: Configuration & Setup, File Processing Fees, and BlackCat Hosting & Data Visualization.

1. PH Configuration & Setup Fees: one time fee, depending on scope, specs and email message volume. Includes processing fees for test set documents, full configuration, testing and installation.
2. File Processing Fees: cover the per file processing fees for Backfile or stored documents. Pricing is per document up to the PH "all you can eat" capped pricing model mark. All you can eat caps the PH processing costs, regardless of ultimate file volume.
3. BlackCat Data Visualization & Hosting: covers BC configuration, data loading from PowerHouse, hosting and access for up to 20 simultaneous users for 3-12 months.

Typical Assumptions

1. Valora leaves the purchase and installation of hardware out of the Phase 1 discussion, as the work is being conducted, managed, hosted and maintained onsite at Valora's Bedford facility or within Client's data center.
2. Valora assumes Client will perform de-duplication of matching files from the stored systems. Our estimates assume a 40% cull rate.
3. Client personnel (and not Valora) will handle Exceptions as they arise.
4. 40% of the emails (incoming & outgoing) will have attachments that also require automated categorization and filing.

PHASE 2: BACKFILE

Objectives

- Provide full-scale analysis, processing, hosting and management of full Backfile and/or historical document population.
- Provide automated indexing, tagging, keywords and other key document metadata to all Backfiles.

- Provide easy-to-use document analysis & data visualization tool for easy, intuitive document population assessment, organization and disposition for full department usage.
- Move newly classified documents from to defined file storage locations, retaining indexing/metadata for future use, and defensibly disposing obsolete and/or irrelevant files.
- Provide robust reporting of backlog/historical data migration activities and progress.
- Run all processes in Valora's high volume environment or onsite at Client's data center.

Phase Two indexes & classifies the rest of the team's accumulated documents (emails, attachments, and any other files to be assessed). The classification evaluates all historical documents for suitability and classification into the full taxonomy of folders and issues. Phase Two operates either at Valora's Bedford facilities, in our high-volume, production environment, or onsite at the Client's data center. Once processing is complete, Valora uploads the documents and all indexing work product to Client's DMS and/or to BlackCat. When hosting the Phase Two documents inside BlackCat, Valora typically also delivers a complete set of Phase Two database load files to client for loading to internal systems.

Plan Steps & Rough Timing

1. Utilize guidance and specifications from Phase 1, including any changes/additions.
2. Valora re-configure PowerHouse and BlackCat to client custom specifications. (Often overlaps with Step 1)
3. Client supplies complete Backfile/historical set(s) of custodian data to Valora or access to stored files in DMS.
4. Valora tests any changed configuration against sample set from new data.
5. Valora begins processing, by custodian or email box priority, moving into high-volume, production mode.
6. Valora populates BlackCat and/or Client DMS with data set & results, with routine, high-volume deliveries.

Timing for Phase Two includes any/all reconfiguration of PowerHouse & BlackCat, prior to high volume file processing. Phase Two duration is typically 3-6 months.

Cost

There are three main fee areas: Re-configuration & Setup, File Processing Fees, and BlackCat Hosting & Data Visualization.

1. PH Re-configuration & Setup Fees: one time fee, depending on scope, specs and email message volume. Includes processing fees for test set documents, full configuration, testing and installation.
2. File Processing Fees: cover the per file processing fees for backlog or stored documents. Pricing is per document up to the PH "all you can eat" capped pricing model mark. All you can eat caps the PH processing costs, regardless of ultimate file volume.
3. BlackCat Data Visualization & Hosting: covers BC configuration, data loading from PowerHouse, hosting and access for up to 20 simultaneous users for 3-12 months.

PHASE 3: DAY FORWARD

Objectives

- Provide full-scale analysis, processing, hosting and management for new documents on an ongoing, permanent basis (Day Forward documents).
- Provide automated indexing, tagging, keywords and other key document metadata to all files.
- Provide easy-to-use document analysis & data visualization tool for easy, intuitive document population assessment, organization and disposition for full department usage.
- Move newly classified documents to defined file storage locations, retaining indexing/metadata for future use, and defensibly disposing obsolete and/or irrelevant files.
- Provide robust reporting of backlog/historical data migration activities and progress.
- Run all processes in Valora's high volume environment or onsite at Client's data center.

Phase Three marks the start of ongoing processing activity, for all files created henceforth. A key component of Phase Three is the transition to live Valora processing and hosting either within Client's offsite data center, or within Valora's offsite data center.

Phase Three optionally includes "Classify at Save Time," a dialog box that pre-classifies new documents live, on the fly, as they are created (saved), and permits the user to edit those attributes, if desired. Other options include: periodic classification "sweeps," staged classifications with manual approvals, and more. Once saved/approved, the documents are now stored in their new folder locations, with proper filename and indexed attributes and keyword tags, including read/write access, and available for searching, data visualization and reporting via BlackCat.

Plan Steps & Rough Timing

1. This is the third and final step of a sequence of email classification and organization activities, and presumes both Phase 1 and Phase 2 steps are complete.
2. Utilize guidance and specifications from Phases 1 & 2, including any changes/additions.
3. Valora re-configure PowerHouse and BlackCat to client custom specifications. (Often overlaps with Step 2)
4. Apply rules-based labeling to each file to determine its ultimate destination and disposition in the new system.
5. Implement either batch-staging workflow or live/near-realtime polling integration. Typically involves custom integration with Client's DMS via API or other means.
6. Integrate with DMS security infrastructure (BlackCat installations only). Ex: Single sign-on integration, retention policy rules integration, etc.
7. Ongoing quarterly "tune ups" to maintain and upgrade software, edit rules, create templates and hold quarterly account management discussions with Client. Valora's Operations Team remains engaged with the Client on an ongoing, quarterly basis.

Timing for Phase Three includes any/all reconfiguration of PowerHouse & BlackCat, prior to ongoing use. Phase Three setup is typically 1-2 months, with quarterly tune-ups thereafter.

Cost

1. PH Re-configuration & Setup Fees: one time fee, depending on scope, specs and email message volume. Includes processing fees for test set documents, full configuration, testing and installation.
2. Ongoing PowerHouse and BlackCat license: typically configured per month or per year, license fees cover all expenses in using, maintaining, tuning and upgrading both products.

BASELINE HARDWARE SPECS

During the needs assessment and configuration phases, PowerHouse needs only a modest amount of hardware. Backfile processing typically presents the greatest load on systems and requires more temporarily. DayForward needs are more modest again, based on the exact spec for that phase.

Valora's software is modular and distributed, which means we take advantage of multiple machines running in parallel, typically virtual machines. For the initial phases, we'd like three VMs (specs. below), which provide sufficient processing power for configuration and testing, and also let us ensure that all inter-machine communication is working properly. During that portion of the project, we'll also do some timing tests to determine exactly how much we'll need to scale up for BackFile conversion. Our estimate at the moment is in the 10-15 VM's range, but it will depend heavily on exactly which parts of the BackFile we're processing on what schedule, as well as what the actual files look like. So our recommendation is to set up three VMs at the outset and figure out the BackFile (peak processing) a little later.

Basic VM Specs:

1. **Controller VM**: 6Gb memory, 150Gb available disk, 2 cores @ 2.5+GHz, Windows 7 Pro 64-bit or Windows Server 2012 R
2. **Small processor VM**: 4Gb memory, 80Gb disk, 2 cores @ 2.5+GHz,Windows 7 Pro 64-bit or Windows Server 2012 R2
3. **Large processor VM**: 8Gb+ memory, 80Gb disk, 2 cores @ 2.5+GHz, Windows 7 Pro 64-bit or Windows Server 2012 R2

The VMs need to be able to find each other on the LAN and Valora Technical Operations will need administrative rights to install software and operate the machines remotely.