



SFL Finds the “Needles” in an Eight-Terabyte “Haystack” of Data and Saves the Client Hundreds of Thousands of Dollars

Situation: A law firm asked us to locate emails sent or received by five custodians that contained keywords agreed upon with opposing counsel and the court. To make the project even more challenging, the client told us that the only available sources of data were tape backups and an email archiving system located halfway around the world in Asia.

The archiving system saved raw, encoded copies of all inbound and outbound email inside of a network appliance. Given the prohibitive cost of restoring hundreds, or perhaps even thousands, of backup tapes to get to the relevant documents, we suggested that we attempt the work against the archived email directly from the network appliance.

Unfortunately, we learned that the archived email was saved in the appliance in a flat structure, regardless of who sent or received it. This meant that all emails circulated among the company’s 20,000 employees were stored in one huge “pile.” It appeared that our hunt for emails by the five designated custodians might grow into a true “needle in the haystack” search.

The archived data from the appliance was exported, compressed and provided to SF Legal, where it tallied a little less than eight terabytes upon delivery. We received 51 .tar archives containing anywhere from tens of thousands to hundreds of thousands of .gz archives – another compressed format. The .gz archives were not named or stored in a way that offered insight into the contents of the file, so the custodian’s email address and date of the message remained a mystery.

Once the .gz archives were expanded, we found that each contained a single email message in a raw format with no file extension. Although all information within the header and body of the email was stored as plain text, the attachments were encoded with a common scheme called “Base64.” We estimated that the entire population totaled approximately 50 million email messages. Aware of the client’s expectation of a rapid turnaround, we set out on our daunting search and delivery mission.

Solution: Upon receipt, SF Legal extracted the content of the 51 .tar archives and saved the individual .gz archives onto lightning-fast hard drive storage arrays. To ramp up performance even more, we configured and deployed a fleet of powerful computers nicknamed the “Octo-Bots” to text index all the sets of .gz files. We then queried these indices to

find the desired custodian names and email addresses in the email headers (To, From, CC, etc.).

After identifying .gz archives containing the desired custodian names, we extracted those messages from their .gz containers. Since the messages were in a raw text format and attachments were Base64 encoded, we wrote a custom program to convert the message to .eml format and restore the attachments to a usable format. We then loaded those messages into our EDD processing tool. The final set of documents was exported to a document review database for final review and production. Everything was completed in about three business weeks.

Benefit: Using conventional EDD culling prices, the budget estimates for a project involving 50 million email messages could have topped seven figures. But with our advanced technology and experienced team, the final cost for our searching and processing efforts was less than 20 percent of that amount.

In addition, after running the data through our custom processes and extracting relevant emails, there were only a few hundred messages of interest to our client. This saved them time and money on document review. In fact, they had such a small set of results that we were able to provide them with a simple Concordance database, obviating the need for web-based, high-capacity hosting. Plus, by working with SF Legal, our client saved hundreds of thousands of dollars for their corporate client, which of course helped strengthen their relationship.

By beginning with the end in mind, SF Legal minimized costs for our client throughout every aspect of this project. Controlling **C**ost, focusing on **P**rocess and minimizing **R**isk for our clients is the cornerstone of our “**CPR**” methodology. To learn more about how SF Legal’s [CPR](http://www.sanfranciscolegal.com) can help you breathe easier, visit www.sanfranciscolegal.com or call us at **415-392-2900**.

