



Data Collection from Google Apps

Gmail, Calendar, Contacts, Documents

Executive Summary

- Google Apps data is becoming a common target for EDD collections.
- The existing tools for collections from the Google cloud are ineffective for mid-size to large collections because of connectivity, speed and verification issues.
- Google Data Collector, a new service from Electronic Legal, is the only solution currently available that provides connectivity, download speed, verification logs, and data security.

Are Google Apps and Gmail really on the horizon for EDD collections?

Certainly in our practice, the answer is yes. Forensic collections of Google App data, particularly of Gmail, are becoming more common. Here are few reasons.

- It's Google. Huge company, the web's defacto search engine, amazing marketing and web presence. An estimated 30 million corporate users in three million companies, as of March 2011, and the growth curve is very steep.
- None of the other mail providers is offering an integrated suite of cloud-based productivity tools that is effectively competing with Google Apps.
- Small to medium size companies appear to be migrating to Google Apps in increasing numbers.
- In particular, start-up companies (often the target for forensic collections in non-compete and IP theft cases) are embracing Google's cloud-based solutions.



- Obviously there is still a ton of webmail sitting in Yahoo, Comcast and Hotmail accounts, but these tend to be used by individuals, not corporate entities.
- Pricing. At fifty dollars per user per year, versus the cost of in-house servers, software and maintenance it takes a lot of users to make in-house infrastructure investments make sense.

What is Google Apps?

Google Apps is a suite of cloud-based applications that provide basic productivity applications. In the best of all worlds, (at least from Google's point of view) they provide alternatives for the most common business computing services. Gmail, Postini, Calendar and Contacts replace Outlook and Exchange. Google Docs replaces Word, Excel, Powerpoint, SharePoint and the corporate file share servers. Although there is no doubt that Microsoft has a huge installed base that is not disappearing anytime soon, Google is making inroads, particularly among smaller, more nimble IT environments.

Our company switched to Google Apps a few months ago, with some limitations. We were quickly able to get rid of our Exchange server and our Outlook clients. Gmail is pretty sophisticated and relatively painless to incorporate. It includes calendar and contacts, is easy to use, administer and access. The biggest drawback is not having offline mail, and there are new services available to address this issue as well. There is no looking back, and certainly pulling the plug on the corporate Exchange server was a wonderful feeling.

Google Docs, on the other hand, is not quite ready for prime time. Lot's of subjective comments in this paragraph so be forewarned. The storage and sharing aspects of Google



Docs is excellent. The ability to simultaneously edit a shared document is fantastic. The editing tools, unfortunately, are pretty weak. Word processing is fair, although nowhere near the capabilities of MS Word. If you are even a mildly sophisticated Excel user the spreadsheet capability is frustrating. And the presentation software is not in the same ballpark as PowerPoint, at least without a significant learning curve. At this point we do most of our mildly complex document work in MS Office, and upload and download it to Google Docs. Having said that, Google Docs is a great repository, and it has already become integral to our business.

What does a corporate Gmail collection look like?

We have had a number of cases that include Gmail collections. This has, of course, led to exploring different technologies, new verification techniques, and a host of new challenges. Let's focus on an example Gmail collection for a ten-person company called (with a grateful nod to years of roadrunner and coyote cartoons) Acme, Inc. Assume the ten Acme employees have an average of 30,000 emails and two gigabytes per user in email storage. Google Apps users tend to have much larger mail boxes than they would have on Exchange server since the default storage allowance is 25 gigabytes per user. The Google search capability makes finding mail in a large data set easy. And, of course, there is no limitation coming from the IT department to keep the mailboxes small or to archive or delete old mail. We regularly encounter single mailboxes above ten gigabytes.

If the company used traditional client-server technologies, and all we wanted was mail, the process would be pretty simple. Grab the Exchange database and supporting files, look for deleted data, search and collect any pst and ost files off the client computers, provide the usual hash checks and custody documents and, for 99% of cases, you are probably ready to move on to processing.



However, if the company is using Gmail, there is no single corporate mail database to collect. There is little or no significant email data on the client computer. All the data is on Google servers, and Google does not provide any method to easily collect or verify the data.

Further, the traditional user mailbox model of a PST file containing emails neatly organized into folders does not apply. Instead of folders, Gmail uses a concept called labels. Labels are virtual folders that create pointers to email, not physical copies of the email. Luckily, there is a label called “All Mail” which contains every email in the user’s Gmail account. That’s the one we’re going to focus on.

Connect to Gmail

Ready, set go. You have the user names and passwords, correct? No? Well that’s a problem.

Obviously, this is a big difference from client/server collections. You can’t even get to the data without the passwords. Even if the collection is approved and scheduled to begin, collecting and managing passwords is a unique problem. But there is a reasonable solution. Google has an administrative log in for corporate accounts. With the administrative log in, you can assign new user passwords, and you can set the account to require a new password the next time the user logs in. Using this tool users can be locked out of accounts during collections (although this is not technically required, collections can be made from live accounts), temporary passwords can be assigned, and the users can be allowed to reset passwords so they are free of further intrusions as soon as the collection is complete. All of this can be accomplished with minimal effort and



participation from corporate IT, assuming they are willing to allow you access to the administrative log in for the duration of the collection.

Connected! Ready to download...

Now the fun really begins. Gmail supports both POP and IMAP, but POP is associated only with the Inbox it is of limited value in a forensic collection. We'll assume that most forensic collections will not attempt to use POP to gather a large mailbox with multiple labels – it's just not practical.

Which leads us to IMAP. We have used and extensively tested at least a half dozen tools and methods. They generally fall into two categories – client interfaces to Gmail such as Outlook or Thunderbird, and tool specific Gmail backup programs.

Using Outlook as an email client

First, let's explore using Outlook as an email client. Once you get past the Outlook IMAP setup and attach it to the Gmail account, you may well encounter the same issues we did. By the way, all these issues apply to using Thunderbird as a client as well. Just substitute mbox for PST.

- **Large PST Corruption.** Outlook has a history of corrupting PST files as they get large. We've all seen it. You can break the collection up into multiple PST files, but this is difficult, manually intensive and hard to verify. Mbox has similar issues.
- **Header vs. Body.** By default, Outlook will only download the message headers in the IMAP connection. Even when you set Outlook to download the entire message, it downloads the headers, then slowly works it's



way through the headers, downloading each message body. You have to manually check the PST from the client software to make sure the message bodies are actually there.

- **No logging.** What you see is what you get. Outlook does no logging, and hence has no verification capability.
- **No automation.** This is not a “connect and just let it run” situation. Connections get lost. Reconnections may or may not be automatic. Duplication of data often occurs. One CCE (Certified Computer Examiner) stated that it took a week to get a twenty gigabyte mailbox through Outlook.
- **Speed (or the lack thereof).** Outlook suffers from the speed issues associated with downloading any data from Gmail, discussed in the section below. Suffice to say, you better not be in a rush.
- **No scalability.** Using an email client to collect one small account is fine. Using it to collect a company’s data, or a large account quickly reveals the limitations.

Using Gmail specific backup tools

We’ve tried all of these we can find: Gmail Backup, Beyond Inbox and a number of others. They are good programs that suffer from some common deficiencies.

- **Backup, not collections.** These programs are designed to provide end users a backup of their Gmail data. They are not designed to complete forensic collections.
- **Minimal logging, no verification.** Some of the Gmail backup utilities create logs. None of them, to our knowledge, are able to produce a report



that meets a reasonable verification standard.

- **No automation.** Like Outlook, this approach is incredibly manually intensive when applied to a corporate collection.
- **Speed.** The same speed discussed above apply.
- **No scalability.** As the collection grows, the tools become less capable.

Speed (Or the Lack Thereof)

Collections from the Google cloud are slow. While we have no specific knowledge of Google's internal connectivity and bandwidth strategies, at least empirically it seems evident that Google throttles any large-scale download. In general, downloads run pretty quickly for the first thousand emails or so, then the download speed takes a nosedive. Collections of a ten gigabyte mailbox using Outlook can take two days to a week.

Google Data Collector

Electronic Legal has developed proprietary software called Google Data Collector. Google Data Collector is part of the eCloudCollect Suite. *At the present time this software is only available as a service from Electronic Legal.*

Google Data Collector collects Google data efficiently and with documented verification. The application allows us to provide the following:

- **Fast collections.** Collect data at up to ten times the speed of the client and backup collection programs.
- **Scalable.** Connect a dozen computers to a single account and run concurrent collections.



- **Collection Reports.** Reports of what data exists in the cloud, what was collected, and exception reporting for noncollectable data.
- **Emails in EML format, documents in MS format.** Data is provided to you in standard formats, metadata (to the extent that Google and MS metadata are compatible) intact.
- **Easily accessible collected data.** The data is encrypted, zipped, and available for download from a secure FTP site.
- **Each zip file contains a verification log.** We can provide a description of methodology if necessary for your report or testimony.
- **Wholesale and retail pricing.** Pricing is by data volume. We offer wholesale pricing to EDD and forensic firms.

Why don't we sell Google Forensic Collector?

We are developing Google Data Collector using a SaaS model. No dates available yet - the modifications from an in-house proprietary tool to a public web software are fairly extensive. For now we will only be offering Google forensic collections as a service.

To discuss your collection (or any other issues you are interested in) please contact:

Michael Horwith 303.209.0911 x. 107
Trent Walton 303.209.0911 x. 102