Factors to Consider in Estimating the Performance of a Harvester Job

Proximity to the data - If all else is equal, we advocate executing Harvester on the system the data resides on, thus limiting network variables (including older hubs, routers, gateways, cables and any concurrency/bandwidth contentions). If network sources or targets must be used, faster speeds will be reached by keeping the logs folder local to reduce the I/O over the network.

Data composition – When copying a large number of small files (i.e. several million 10KB text files), speed will be limited by disk I/O. When copying a small number of large files (i.e. two 4TB ISO files) speed will be determined mostly by the CPU speed of the machine running Harvester.

Hardware - Aspects that may affect performance include:

- Hard Drive –
    - Type - SSD, SATA, PATA, SCSI (daisy chained or not) IDE, etc.
    - Age - 5400 vs. 7200 vs. 15000 rpm, number of bad clusters, precision of read/write heads, limited buffer cache impacting Disk I/O performance.
    - Seek time - files and file fragments not contiguous, haven't defragmented in a long time, anomalies in the MFT, etc.
    - Configuration - RAID vs. simple volumes.
    - Physical disk and channel separation. Copying from one physical disk to another similar physical disk on the same system is faster than copying to a different location on the same disk.
- RAM - amount and type (i.e. DDR2, DDR3, DDR4)
- Motherboard - slow front side (or EV6) bus or older north bridge chip set vs. HyperTransport, Direct Media Interface, availability of high end Graphics Processing Unit for peak load times, etc.
- Ports - USB 2.0 vs 3.0/3.1, FireWire, older NIC card not supporting full duplex, loose connections, older CAT-5 cables, bent fiber cables, etc.
- CPU – Single-core vs. multi-core, limited L-1 and L-2 on die cache, older architecture and instruction sets, etc.

Software - Full Disk encryption schemes (BitLocker, CheckPoint, Sophos, McAfee), aggressive anti-virus software, virtualization, emulators, hypervisors, middleware managing SAN storage, etc.

Target/Log Paths – Keep everything local if possible, but at the very least try to keep the logs path local. Also, end your logs path with the date/time variable "[DateTime]" to ensure that rerunning the same job will not cause conflicting database errors.

Large Collections – For large collections (i.e. multiple TB of data), try to break source down into multiple smaller sources to be collected by multiple machines running Harvester to the different segments. This will reduce the overall time it takes to collect the full source.

Abbreviated List

- Get close to the data. Try to not collect over any network links.
- Run Harvester installed on the system rather than from an external drive.
- Roll off any full disk encryption prior to running Harvester.
- If there's more than one host system available, use the one with…
    - Newer hard drives, motherboard and current OS
    - More RAM
    - Multi core processors
    - Higher front side bus clock speed