Environment sizing guidelines

To size an environment precisely, you must understand the factors such as harvest frequency, complexity of the source, and use case scenarios that drive application use and action execution.

The general design guidelines for IBM StoredIQ are as follows:

- For data servers of the type DataServer - Classic:

  - One data server for up to 30 TBs of data (which can vary based on the number of volumes, objects per volume, and object types).
  - Up to 500 volumes per data server.

    - ⓘ **Tip:** When you're sizing an environment that includes Sharepoint data sources, keep in mind that volumes must be defined at Sharepoint site collection level, not the Sharepoint server level.
  - Up to 150 million objects per data server.
- One gateway per 50 data servers.
- One application server.
- NFS is slightly faster than CIFS for metadata only, but assume CIFS/NFS even for this exercise.
- Full-content processing of file (for example, .ZIP, .RAR, .GZ) and email archive (.PST, .NSF, .EMX) processing are slower as items must be extracted from the archives. If there is a significant number of these files in the file system and they are not excluded from content processing, the full-content processing rate can be too high. Until you have an initial index of the file system, you do not know how to weigh full-content processing of archives.
- An object/time metric is appropriate for metadata only NOT computing a hash, membership in the National Institute of Standards and Technology (NIST) or enumerating objects that are contained in archives. Converting it to a bytes/time metric is a function of the average object size and might vary tremendously. An average object size of 250 KB was used for the metric that is provided earlier.
- A bytes/time metric is appropriate for metadata-only computing a hash and full-content processing. The object per second rate can vary tremendously depending on the object type and sizes encountered. For example, processing an email or file archive is much more expensive than a PDF document.
- Metadata-only not computing a hash, membership in the NIST list, or enumerating objects that are contained in archives is requesting only the file-attribute information from the NAS. Individual files are not opened and read. The processing rate is high, but that does not translate into a large amount of data that traverses a network between the NAS and data server. The bytes/time rate does not translate into bytes served by the NAS and sent over the network.
- Metadata-only computing a hash, membership in the NIST list, or enumerating objects that are contained in archives opens and reads the contents of each file. The content of all requested files traverses the network between the NAS and data server. The maximum load that the data server can place on a NAS is metadata-only processing. It requires all file content to be read to compute a hash or enumerate objects that are contained in archives. The bytes/time rate translates into bytes served up by the NAS and network traffic that must be considered.
- Full-content processing opens and reads the contents of each file to extract all text. The content of all requested files traverses the network between the NAS and data server. The processing time to enumerate archives, extract text, index words, and extract entities on the data server reduces the rate that data is requested from a NAS compared to metadata-only with full hash. The bytes/time rate translates into bytes served up by the NAS and network traffic that must be considered.
- The interrogator process count on the data server for "metadata only not reading all content indexing" is set to eight for optimal performance.
- The interrogator process count for all other processing that involves reading all content default setting is four per data server.
- The interrogator count can be viewed as the number of client connections that are made to a data source actively requesting data. It is important for capacity planning for the data source.

- The data servers are assumed to be "network close" to the NAS data sources. Network latency under 10 ms with at least 1000 Mbps bandwidth is assumed (connected through local area network). The data servers need a low latency high-bandwidth connection to a NAS data source for acceptable indexing performance.
- The gateway and application stack can be located remotely from the data servers. Network connections with latency greater than 10 ms and bandwidth of at least 2+ Mbps are acceptable.