Glossary

job:

Processing unit representing a single document or entity within a crawl. May hold metadata.

seed:

[*Config Entity*] Crawl starting point, generally a URL relative to a particular server.

connection:

[*Config Entity*] Details regarding how to connect to a given repository server or service.

credential:

[*Config Entity*] Authentication details to access a given repository server or service, generally Username/password or AccessKey/SecretKey pairs.

connector instance:

[*Config Entity*] Base config entity determining type of source repository, and common crawling behavior

workflow:

[*Config Entity*] Set of rules (grouped by workflow event) to be executed sequentially for every given *job* being processed. See Workflows

workflow event:

A virtual set of rules to be executed sequentially that lives inside a workflow object.

throttle policy:

[*Config Entity*] Set of properties determining how often should *jobs* be processed, can be assigned in *credentials*, *connections* or *seeds.* See Throttling Policies

routing policy:

[*Config Entity*] Property that determines where a *job* should be processed (in which *worker node*). See Routing Policies

schedule:

[*Config Entity*] Set of properties that determines how frequently should a crawl start, can be associated with a set of *seeds*. See Schedules

sequence schedule:

[*Config Entity*] Special type of schedule that determines the sequence of crawl starts (after which crawls should other crawls start). See

identity crawl:

[*Crawl type*] Crawl that connects to Identity directories (Ldap, Azure Directory) to cache and process the identities as *jobs*

full crawl:

[*Crawl type*] Crawl that retrieves all documents starting at a given point *(seed)*

incremental crawl:

[*Crawl type*] Crawl that retrieves only the changed documents relative to previous crawls