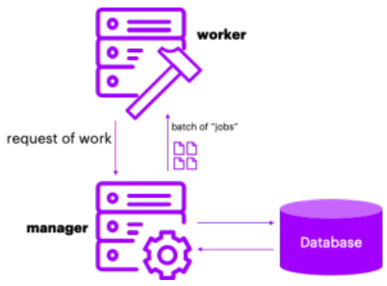
Aspire 5.0 Architecture

Aspire Manager/Worker Architecture

Aspire 5.0 introduces two types of nodes: **Manager** and **Worker.**



A **Manager** is responsible for coordinating the execution of "jobs" from any given crawl and the crawl state, it prepares batches of jobs for eventual assigning to **Worker** nodes.

- There is an automatically elected **main Manager** who coordinates which manager will get to manage each crawl
  - it also takes appropriate actions when either a **Manager** or **Worker** node is detected to be down.

A **Worker** is responsible for processing batch of "jobs" obtained from the Manager nodes.

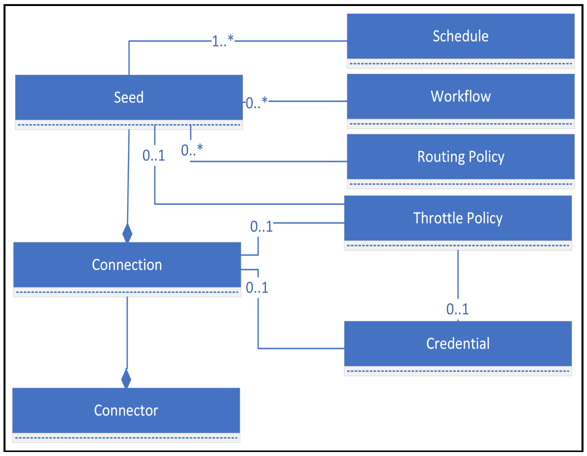
- Also executes all rules inside any workflows configured for the associated job crawls.
- Fetching of content from repositories
- Content and metadata modification/extraction
- Indexing of documents with Publishers

Crawl Configuration

Crawls are now configured in separate entities, which allows for maximum re-usability.

- **Connector instance**
  - Common connector behavior, define number of threads, queue sizes, text extraction capabilities, etc.
- **Credential**
  - To authenticate to a specific repository
  - Authentication Type, user/password, access/secret keys, etc
- **Connection**
  - Properties related to how to connect to the repository
  - Server IP/host/port
  - Connection properties (timeouts, concurrency, etc)
  - Can be associated with 1 credential (if the connector requires credentials to be set).
  - Must be associated with 1 connector instance
- **Workflow**
  - Sequence of rules to be executed for each document
- **Seed**
  - Starting point of a single crawl to execute
  - Can be associated with 0 or more workflows
  - Can be associated with 0 or more routing policies
  - Can be associated with 0 or 1 throttle policy
  - Can be part of 0 or more schedules
  - Must be associated with 1 connection
- **Schedule**
  - Define how often to execute crawls for a set of seeds
  - Define sequence of crawls (chained schedules)
    - For example: start seeds [d, e, f] (chained schedule#2) after seeds [a, b, c] (schedule#1) are done.
- **Throttle and Routing Policies**
  - How often and where should documents be processed
    - Allowing for geo-located job processing
  - Routing policies can be associated with seeds only
  - Throttle policies can be attached to seeds, connections and credentials

Aspire 5.0 config entity model


Schedule
1..*
Seed
0..*
Workflow
0..1
0..*
Routing Policy
Throttle Policy
0..1
Connection
0..1
0..1
Credential
Connector