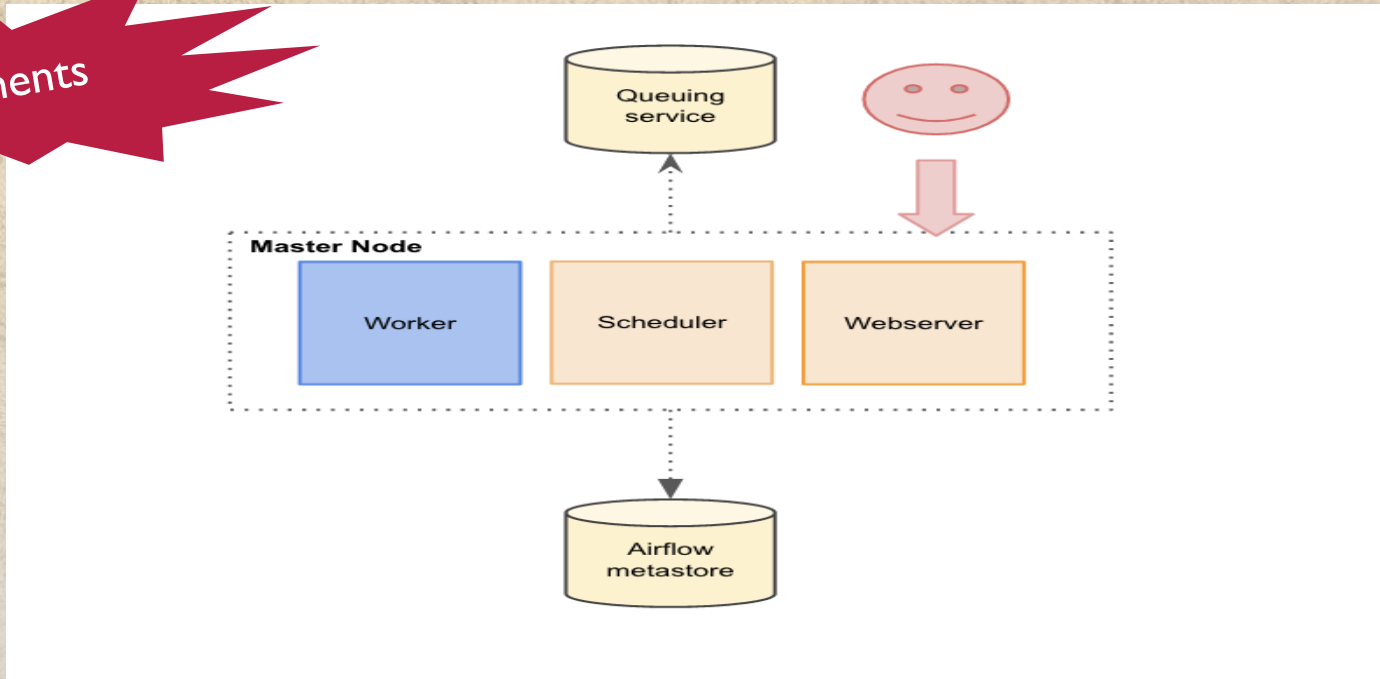
APACHE AIRFLOW ARCHITECTURE OVERVIEW

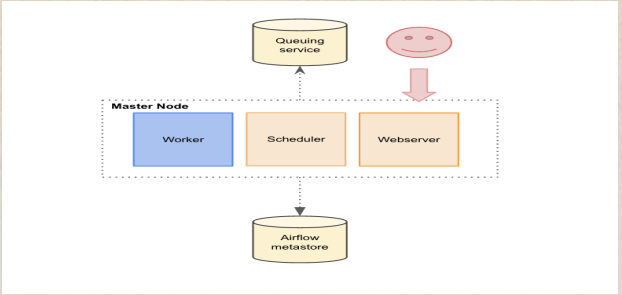
Airflow components

components



Airflow components

- **Webserver:** The airflow webserver accepts HTTP requests and allows the user to interact with it. It provides the ability to act on the DAG status (pause, unpause, trigger).
- **Scheduler:**
  - The Airflow scheduler monitors DAGs. It triggers the task instances whose dependencies have been met.
  - It monitors and stays in synchronization with a folder for all DAG objects, and periodically inspects tasks to see if they can be triggered.

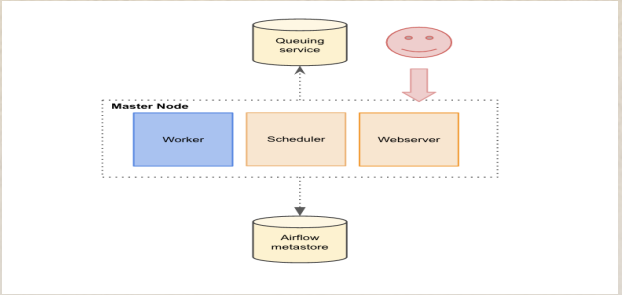


Airflow components

- Worker : Airflow workers are daemons that actually execute the logic of tasks. They manage one to many CeleryD processes to execute the desired tasks of a particular DAG.

  Note :

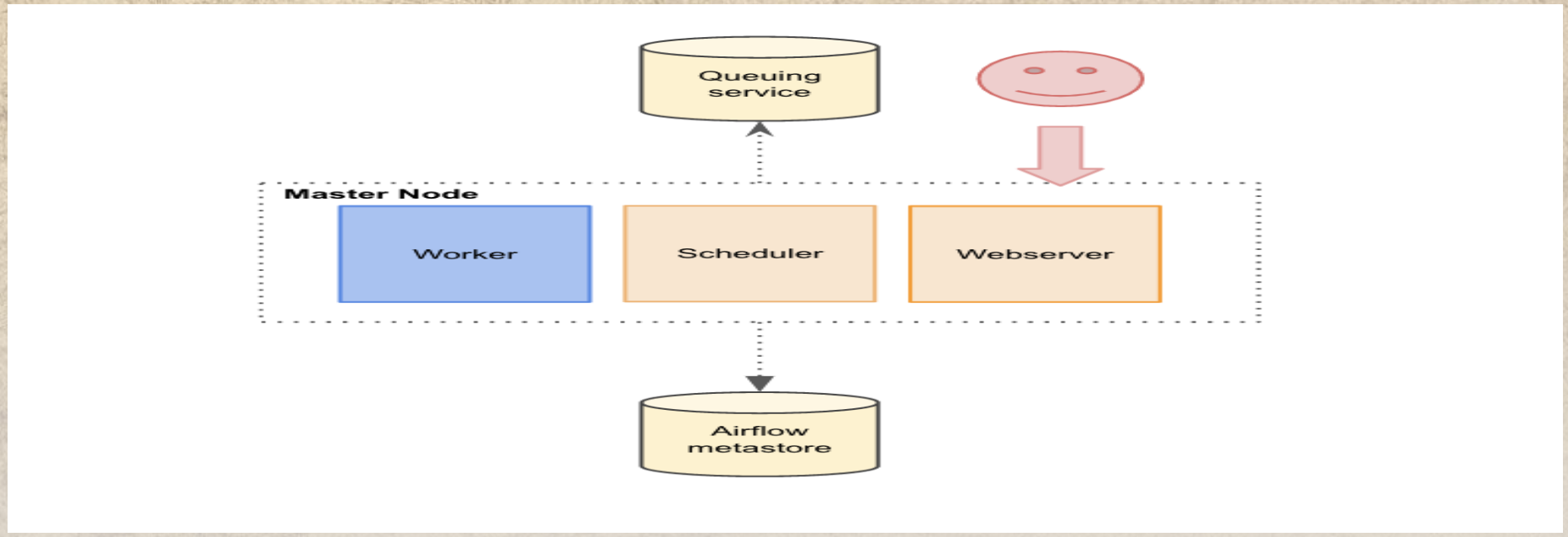
  - daemon is a background process
  - Celery is an asynchronous queue based on distributed message passing.



How they work

- Airflow daemons don't need to register with each other and don't need to know about each other. They all take care of a specific task and when they are all running, everything works as expected.
- The scheduler periodically polls to see if any DAGs which are registered need to be executed. If a specific DAG needs to be triggered, then the scheduler creates a new DagRun instance in the Metastore and starts to trigger the individual tasks in the DAG.
- The scheduler will do that by pushing messages into the queuing service. A message contains information about the task to execute (DAG_id, task_id..) and what function needs to be performed

REVISIT…

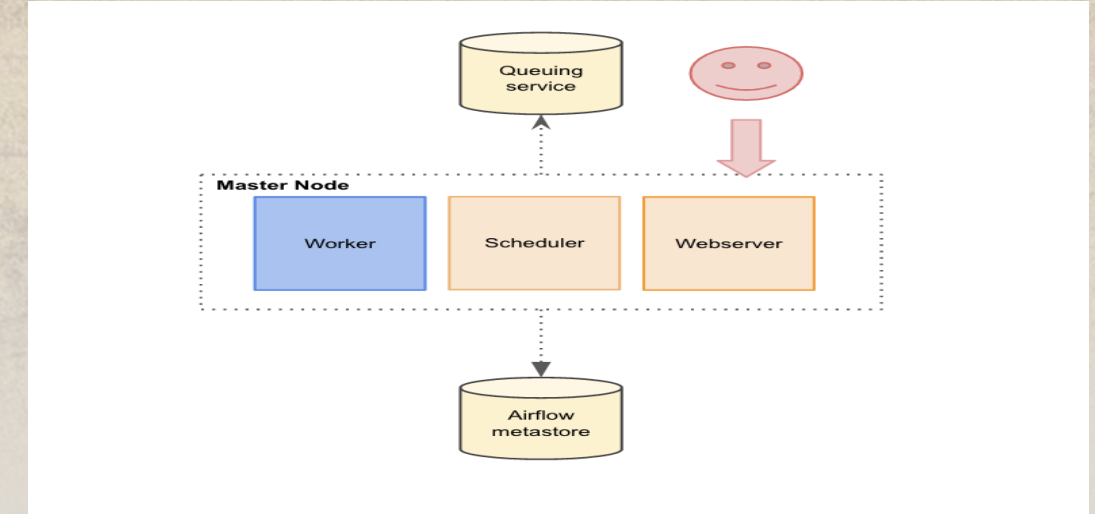

Queuing service
Master Node
Worker
Scheduler
Webserver
Airflow metastore

How they work

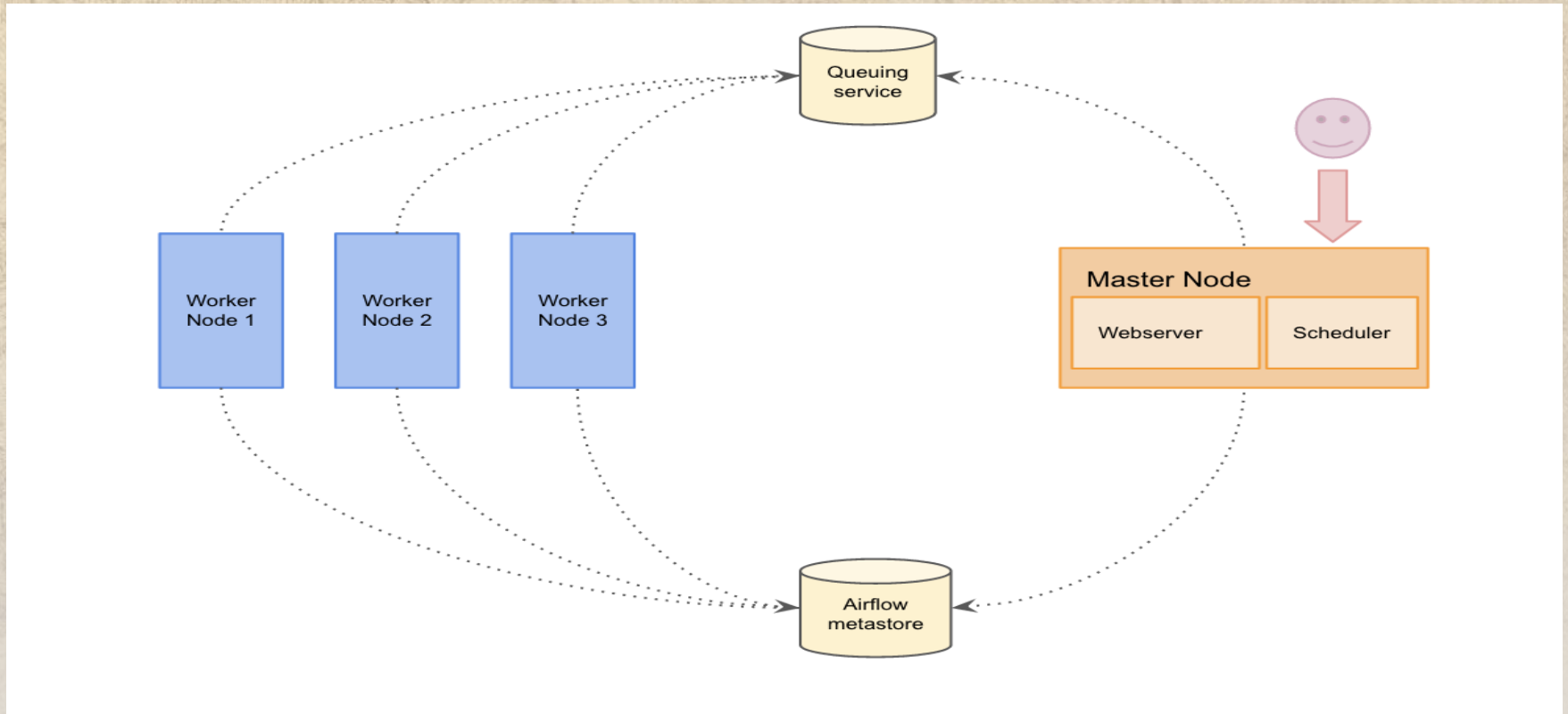
- Celeryd processes, controlled by workers, periodically pull from the queuing service. When a celeryd process pulls a task message, it updates the task instance in the metastore to a running state and begins executing the code provided.
- When the task ends (in a success or fail state) it updates the state of the task in User interface.

Single **vs** multi node architecture

- **Single-node architecture :**
  - In a single-node architecture all components are on the same node.
  - To use a single node architecture, Airflow has to be configured with the **LocalExecutor** mode.



Multi node architecture



Multi node architecture

- In a multi node architecture daemons are spread in different machines.
- To use this architecture, Airflow has to be configure with the **Celery Executor** mode.
- Airflow uses it to execute several tasks concurrently on several workers server using multiprocessing. This mode allows to scale up the Airflow cluster really easily by adding new workers.

Choose multinode for high availability and for scaling purpose