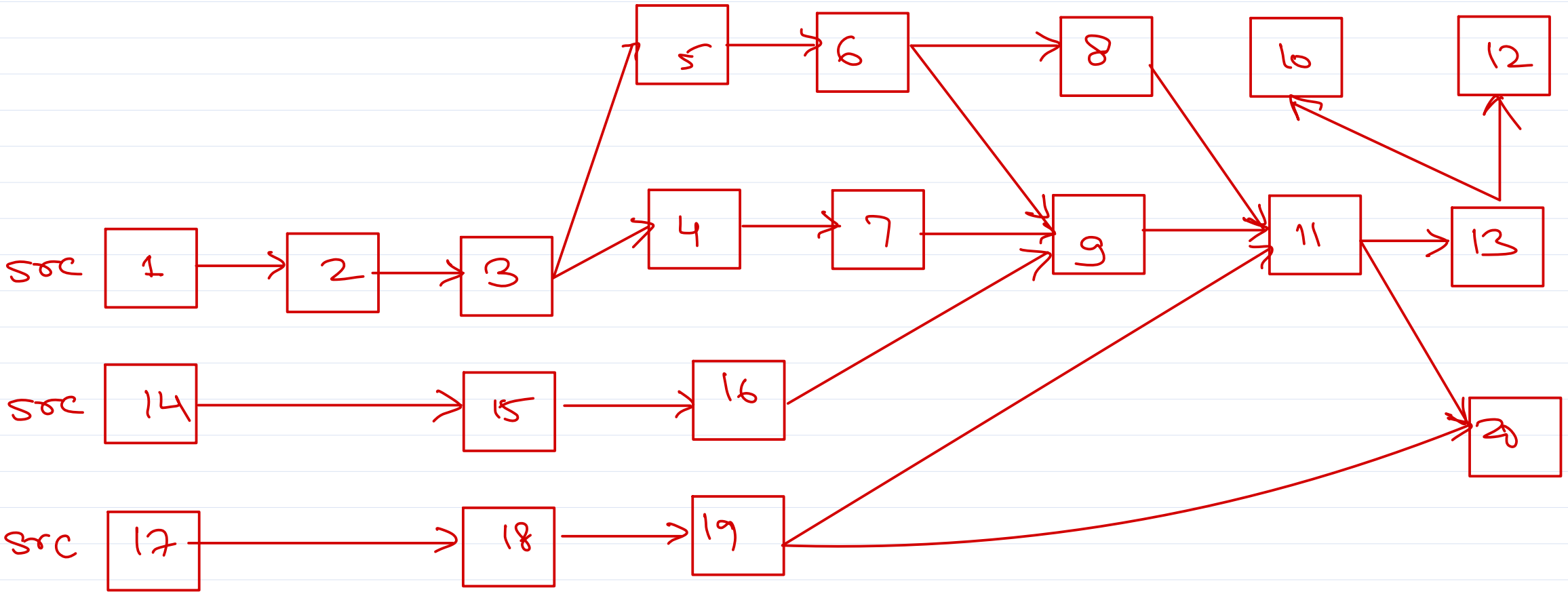
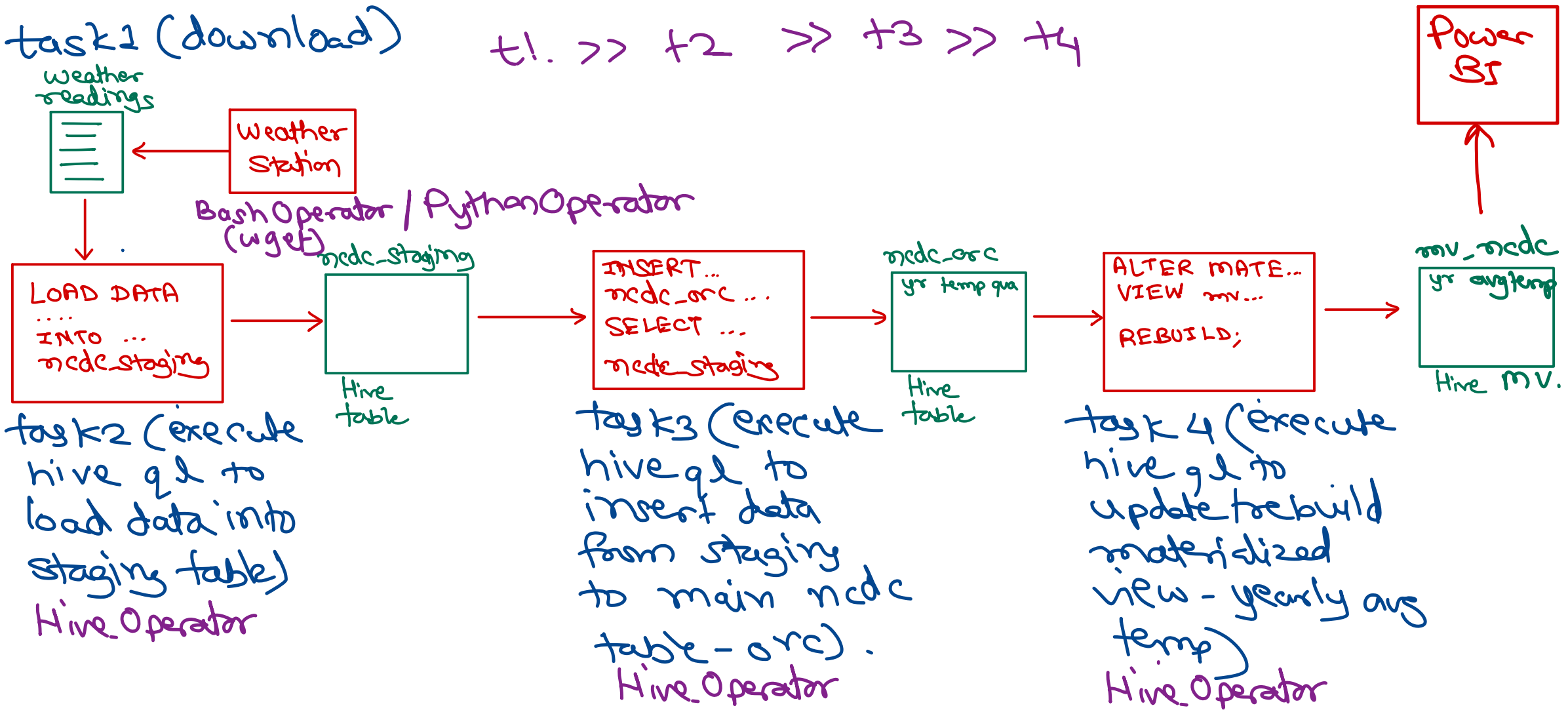
Big Data Technologies

Trainer: Mr. Nilesh Ghule.

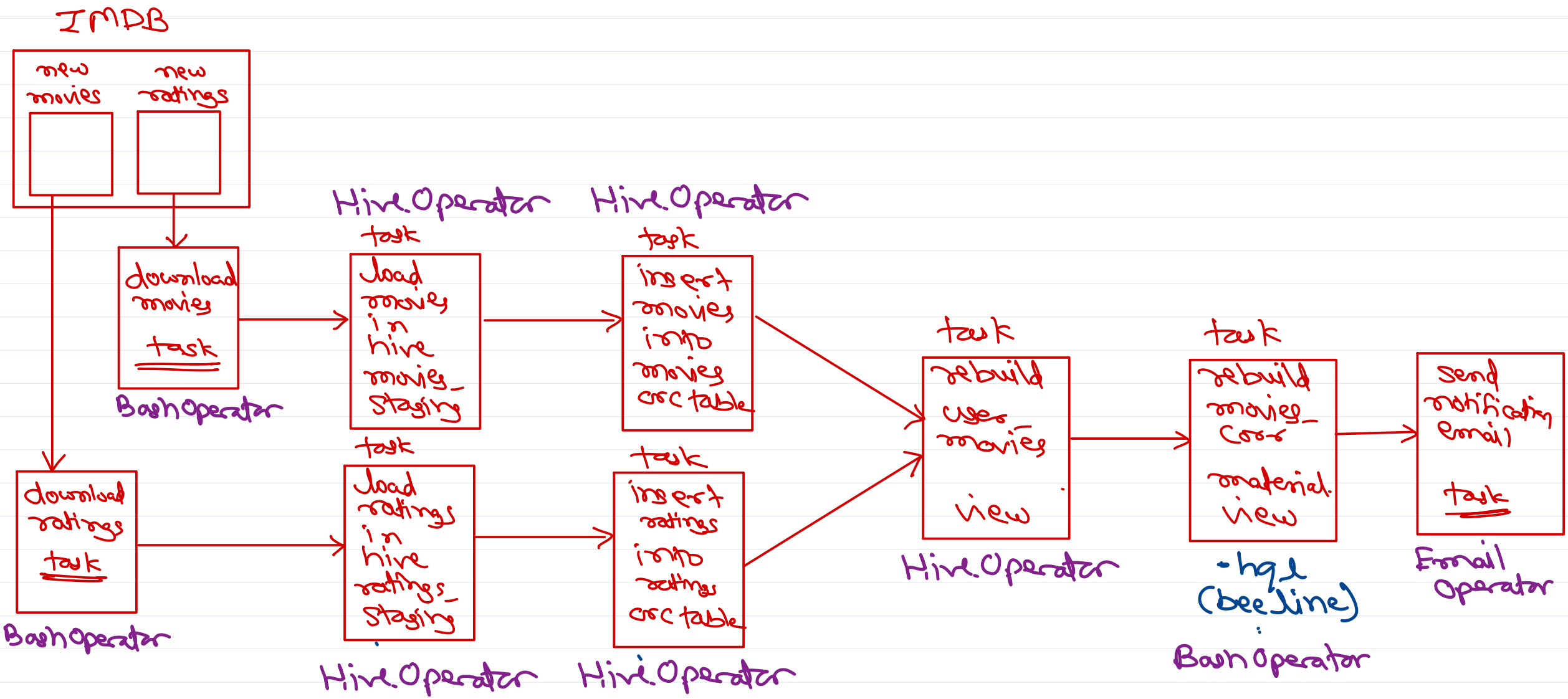
Complex DAGs



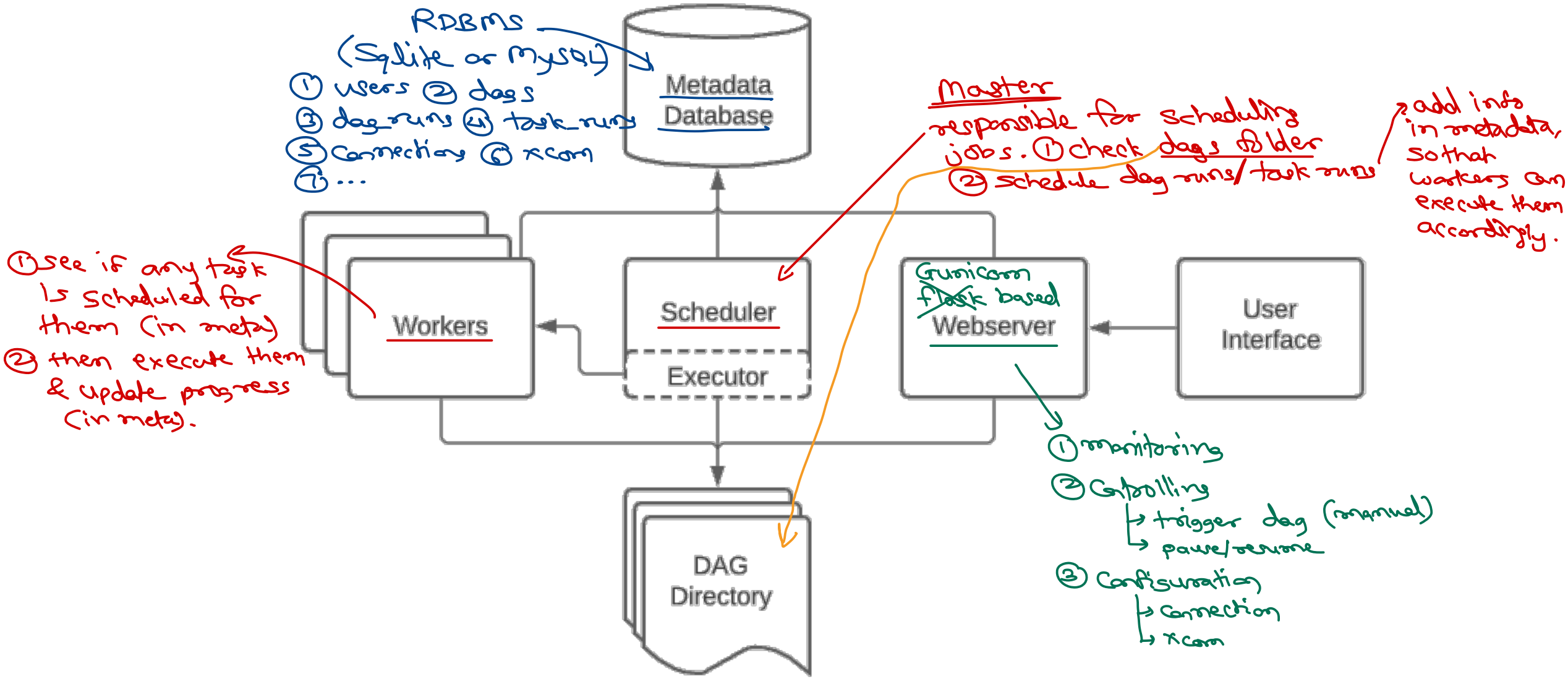
DAG


task1 (download)
t1. >> t2 >> t3 >> t4
weather readings
Weather Station
BashOperator / PythonOperator (wget)
Power BI
LOAD DATA .... INTO ... ncdc_staging
ncdc_staging
Hive table
INSERT... ncdc_orc... SELECT ... ncdc_staging
ncdc_orc
yr temp qua
Hive table
ALTER MATE... VIEW mv... REBUILD;
mv_ncdc
yr avgtemp
Hive MV.
task2 (execute hive ql to load data into staging table) Hive_Operator
task3 (execute hiveql to insert data from staging to main ncdc table - orc). Hive_Operator
task 4 (execute hive ql to update to build materialized view - yearly avg temp) Hive_Operator

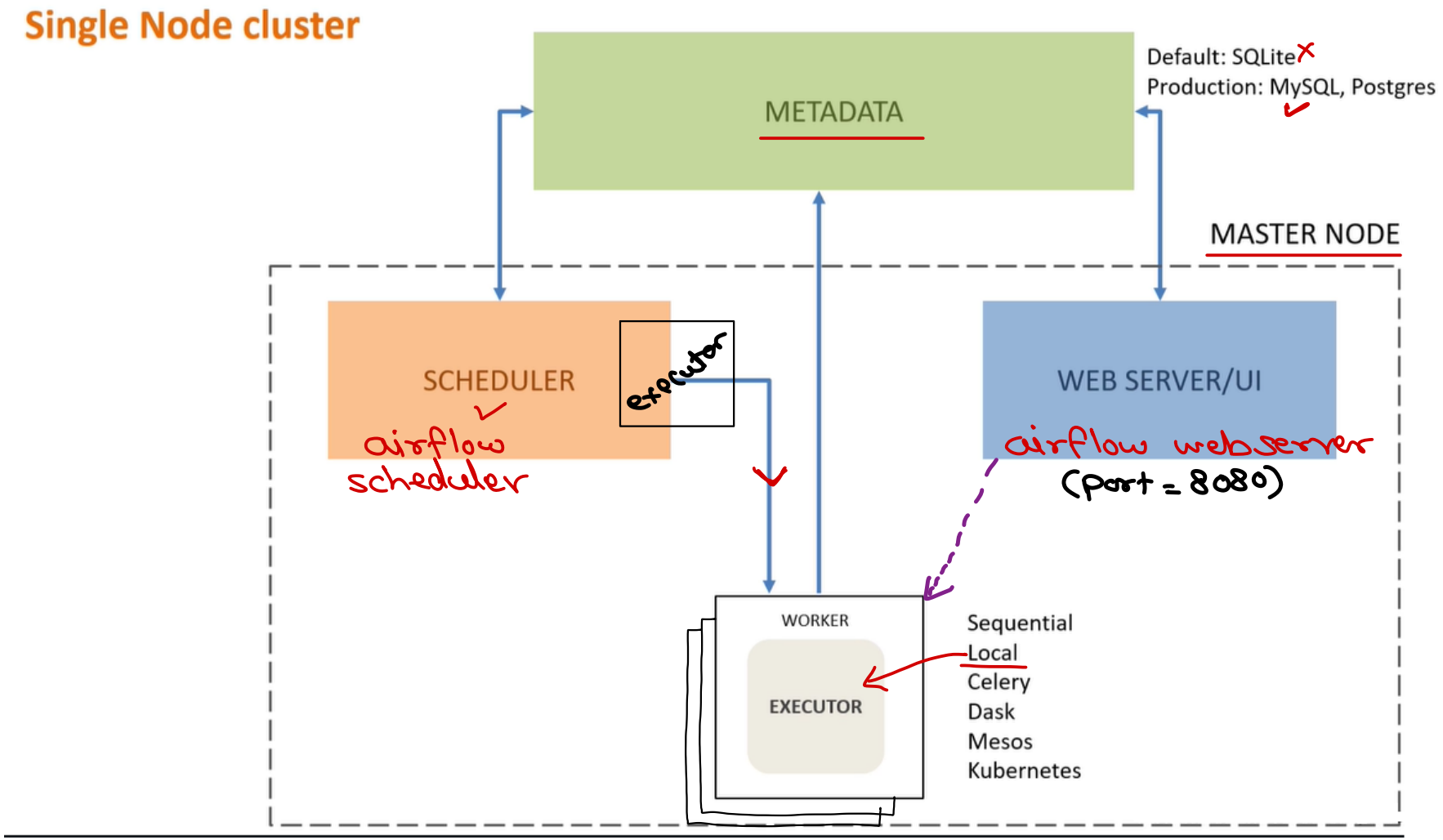
Movie Recommendation Workflow



Airflow Architecture

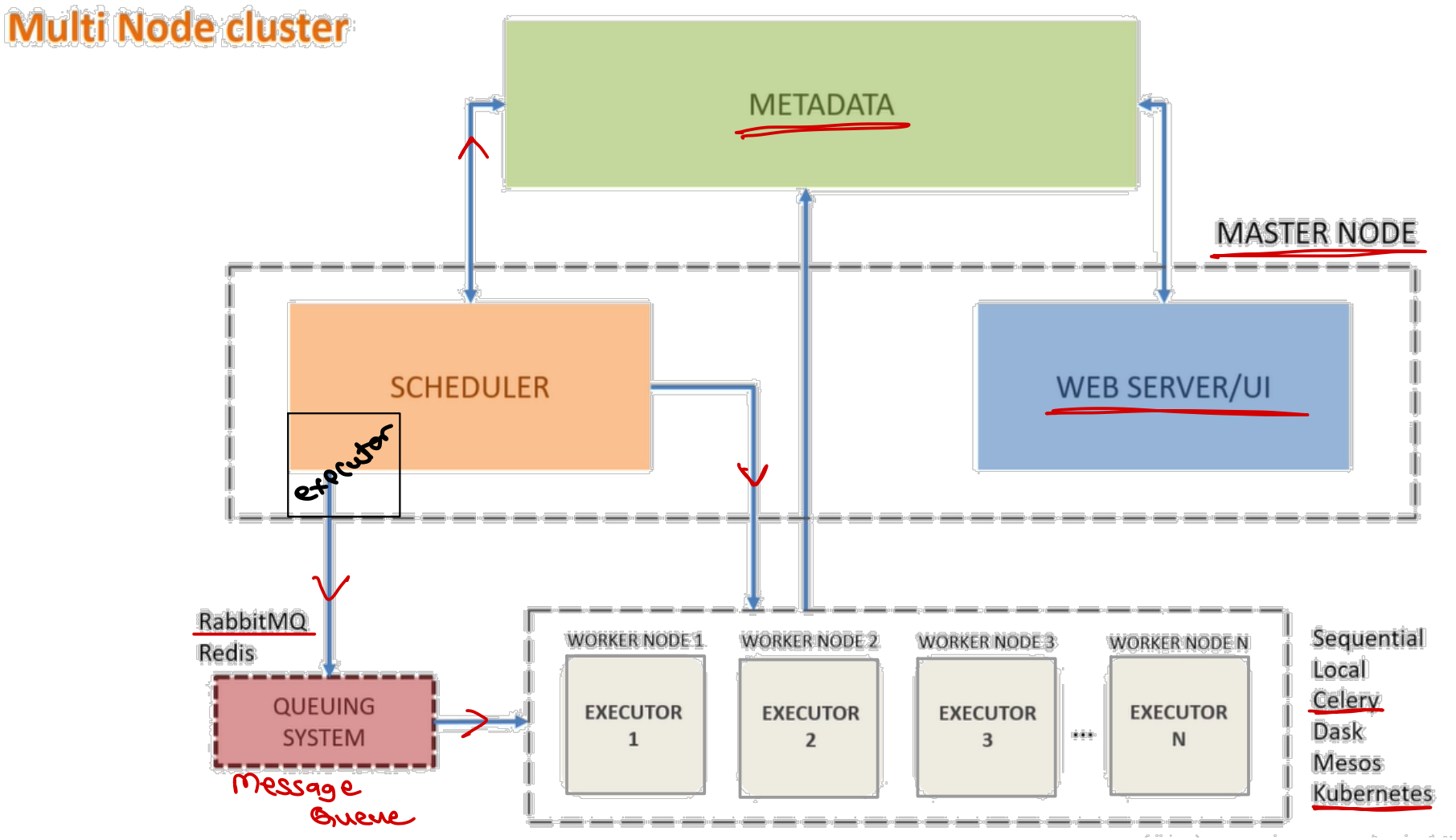


Airflow – Single node installation



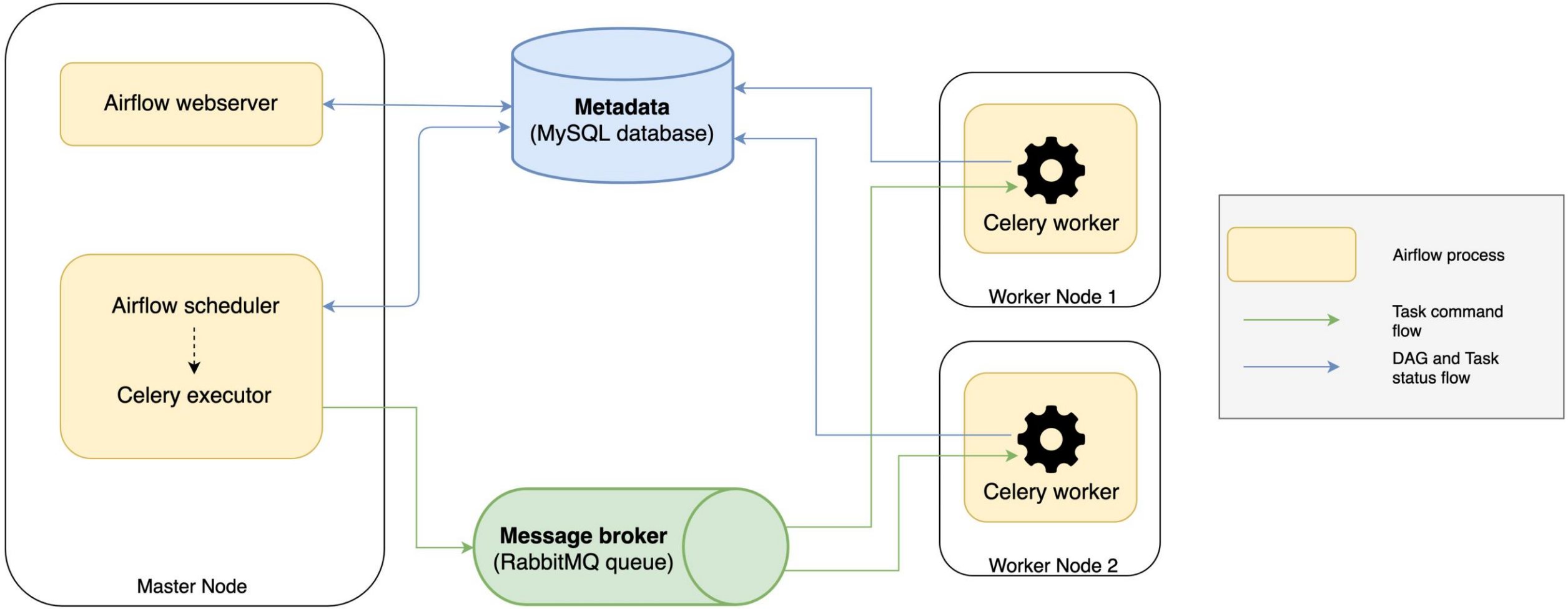
- Single worker node.
- Not scalable (max all resources available on the system).
- Quite fast for limited number of DAGs and data size.
- Use local executor.
- Direct communication between scheduler and ~~executor~~ workers

Airflow – Multi node installation

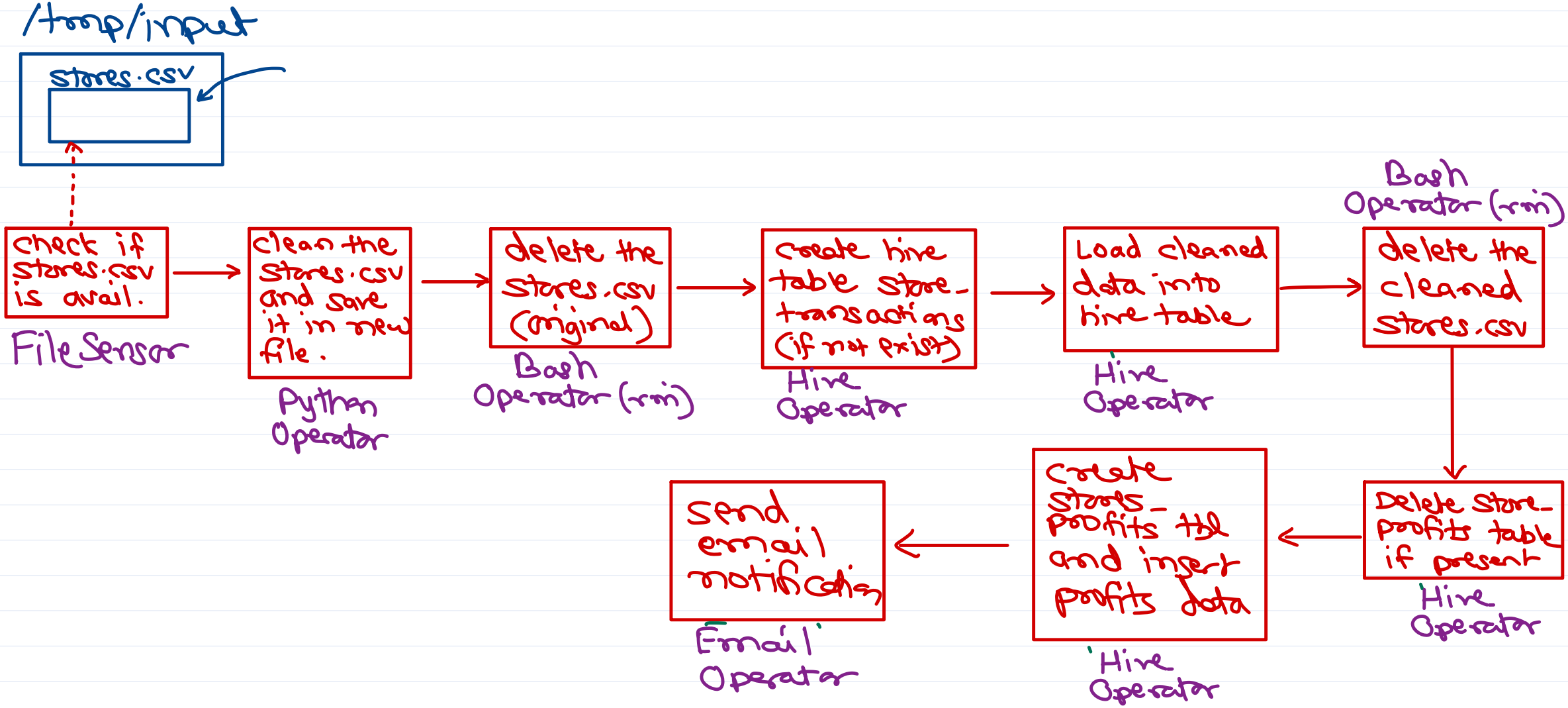


- Master node runs scheduler and web-server.
- Multiple worker nodes - running executors/worker
- Highly scalable for huge data and many nodes.
- Recommended executor is Celery.
- Scheduler and workers communicate with external queue system like Rabbit MQ / Redis.

Airflow – Multi node installation



Data ingestion and Reporting workflow



Thank you!

Nilesh Ghule <nilesh@sunbeaminfo.com>