Why Apache Airfllow Sucks

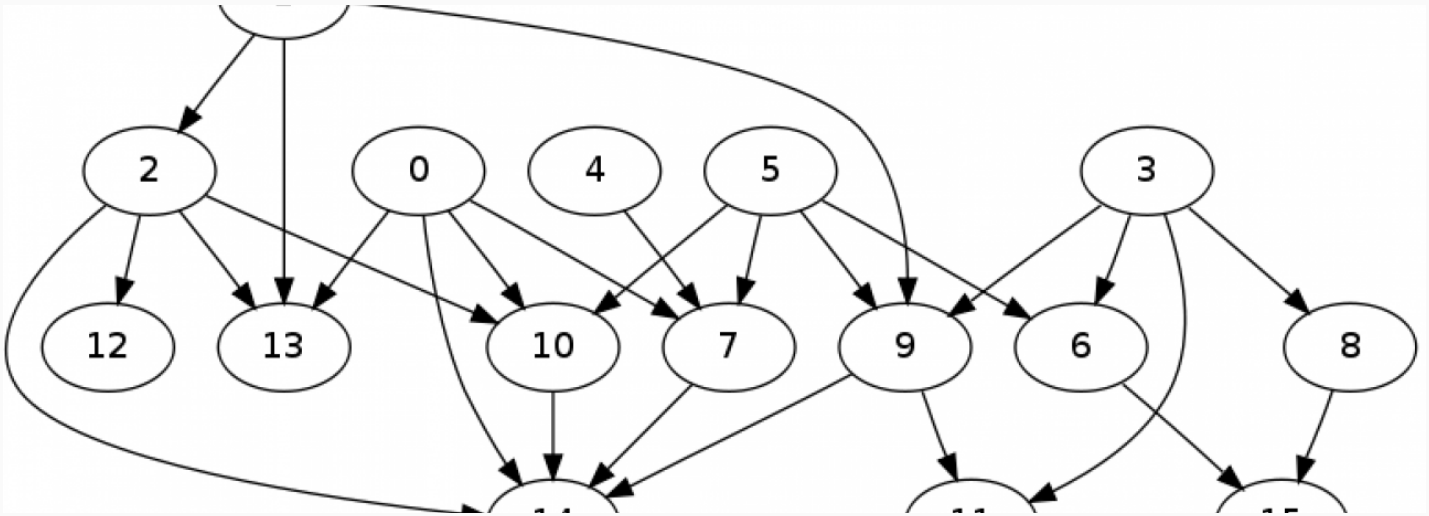
Слава Могилевский

5 октября 2017 г.

Provectus

Введение: для чего нужен Airflow

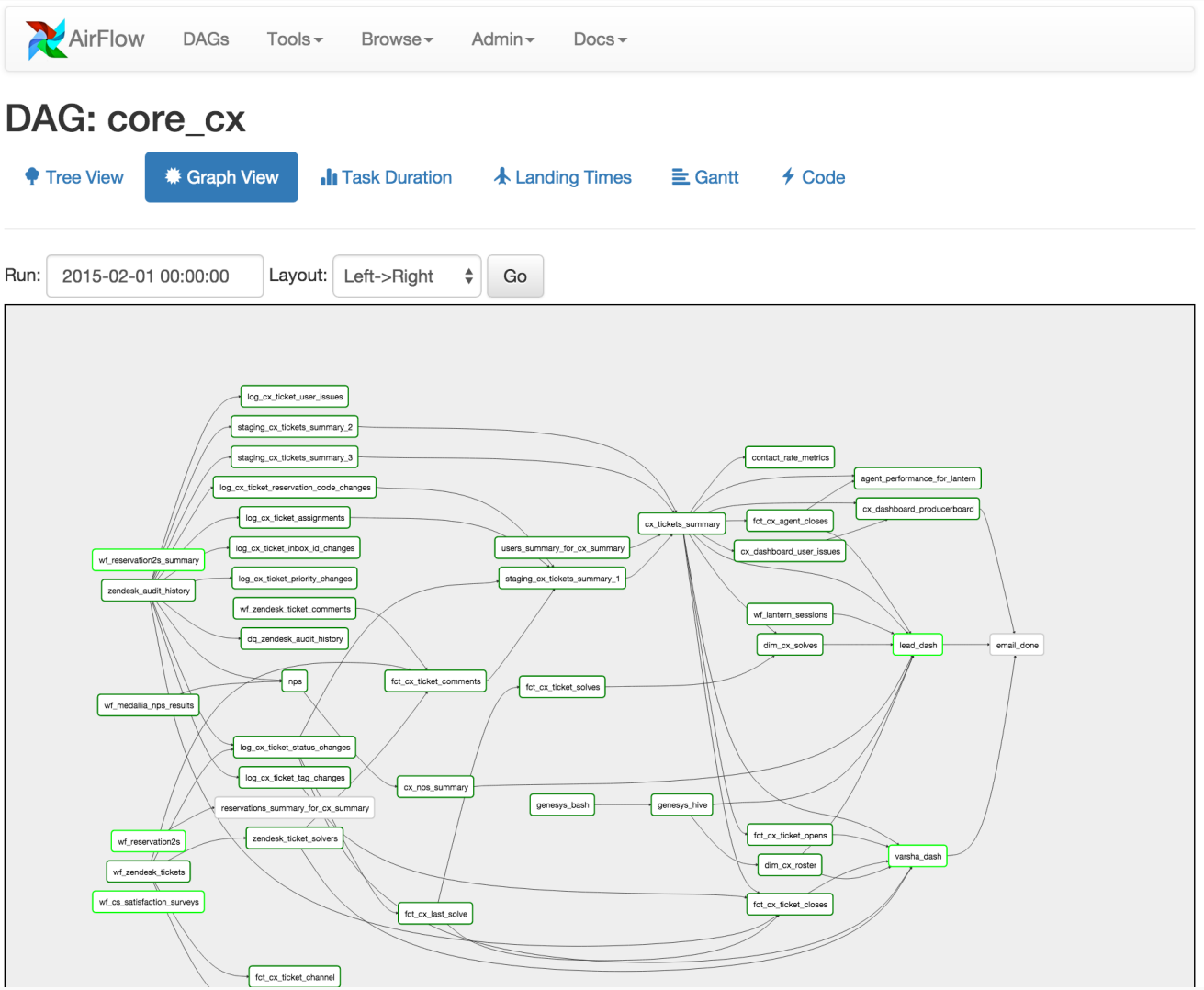
- запустить задачи (hive-запрос, spark, запрос в mist?) по расписанию
- у задачи есть зависимости от других (DAG)
- мониторинг, логи, и т.д



Airflow don't suck sometimes

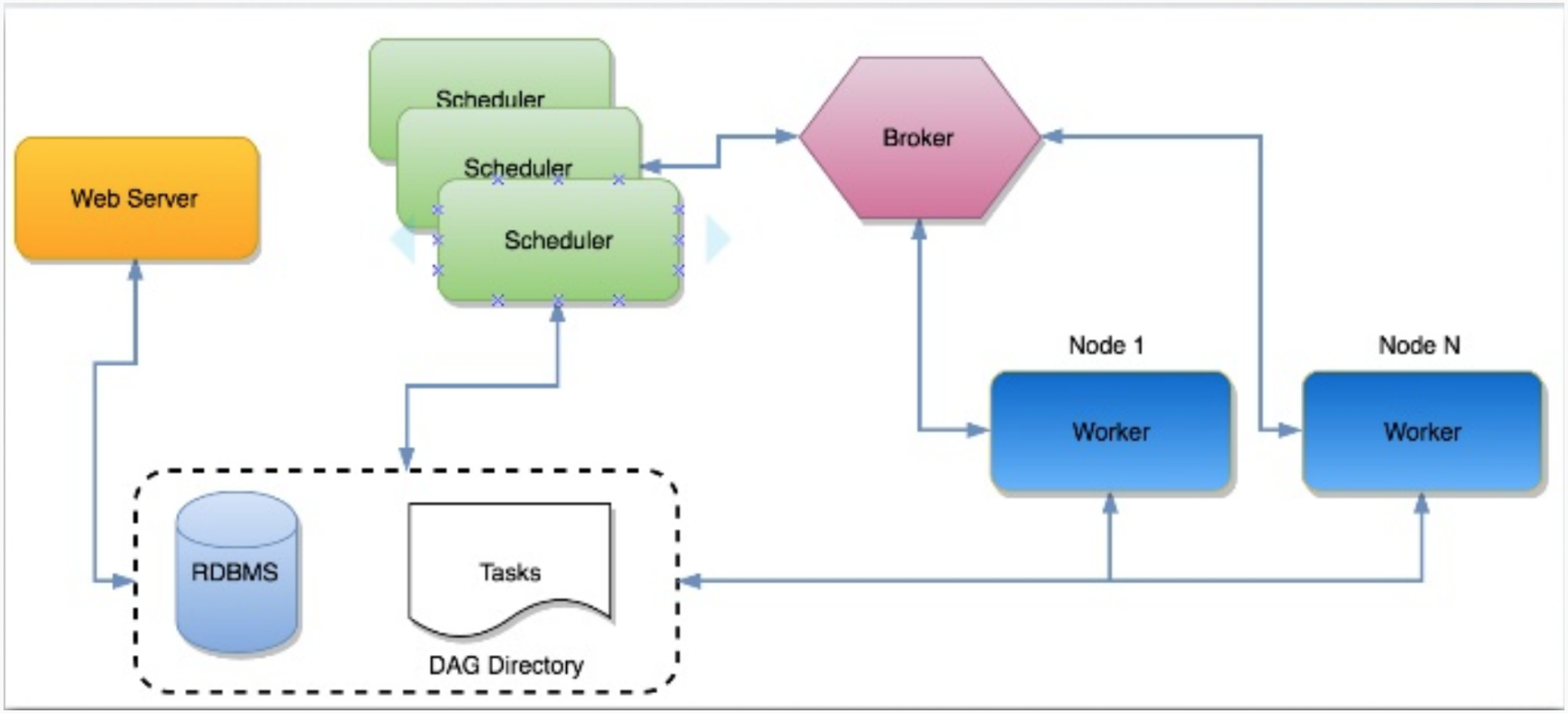
- Динамика: dynamic pipeline generation
- Гибкость: parameterizing scripts by Jinja templating engine
- Масшатибируемость: ready to scale to infinity

Airflow don’t suck sometimes



Удобный UI

Архитектура Airflow



Минусы Airflow

Обновление DAG - это боль:

- Нужно каждый раз рестартить: UI, scheduler
- Обновленный DagBag как-то нужно расшарить между UI, scheduler и Worker'ами
  - NFS - привет, race-conditions
  - Делаем копии DagBag везде, куда нужно - проблемы с синхронизации
- Удаление тасок - часто их еще нужно удалять ручками с БД

Минусы Airflow

Для распределенных worker'ов используется только Celery

А Celery - это тоже боль

Невозможно понять, что происходит с воркерами:

- какая таска выполняется
- почему в airflow висит таска, а в celery все воркеры свободны!?

Минусы Airflow

Логи - это боль

- Логи хранятся у worker'а - они пропадают вместе с ним
- Есть опция хранить в логи S3 - не всегда это нужно
- Хочется иметь живую консоль, как у Jenkins

Минусы Airflow

Системные зависимости в worker'е - это боль

Для каждого нового типа тасок (hive query, spark query) необходимо:

- пересобрать образ для воркера для новой зависимости
- передеплоить новый образ на все воркеры
- ничего при этом не поломать

Альтернатива

Workflow as a Code vs. Workflow as a Infrastructure

Все должно быть взаимозаменяемым (плагины)

Таска = собираем зависимости + go for it!

Еще один awesome-list

`https://github.com/meirwah/awesome-workflow-engines` - список движков для построения и запуска workflow (Jenkins'a в нем нет ;))

- Airflow
- Luigi
- Pinball
- ...

Внимание!!!

Спасибо за внимание