tech

Airflow в кровавом энтерпрайзе

18.09.202

обо мне

Илья Савин

Инженер ИТ инфраструктуры

tech

что такое Apache Airflow

Это открытое программное обеспечение для создания, выполнения, мониторинга и оркестровки потоков операций по обработке данных. (c) wikipedia

tech

Основные компоненты apache airflow

Scheduler

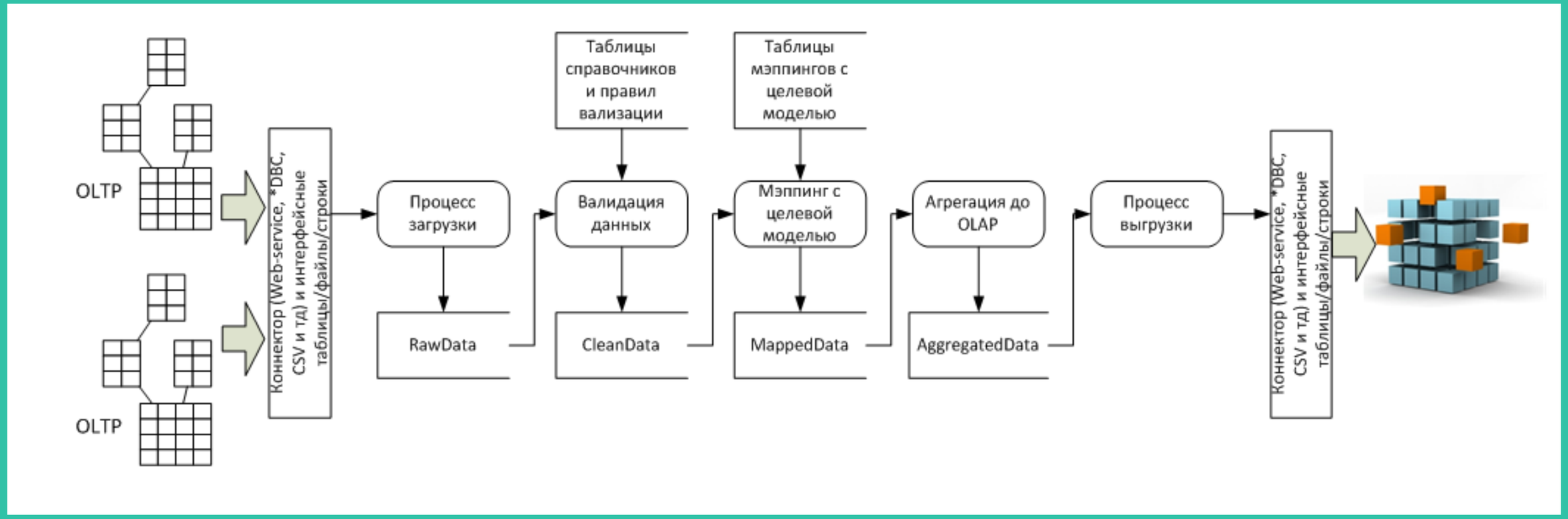
Triggerer

Worker

Web-server

что такое ETL

Extract, Transform, Load

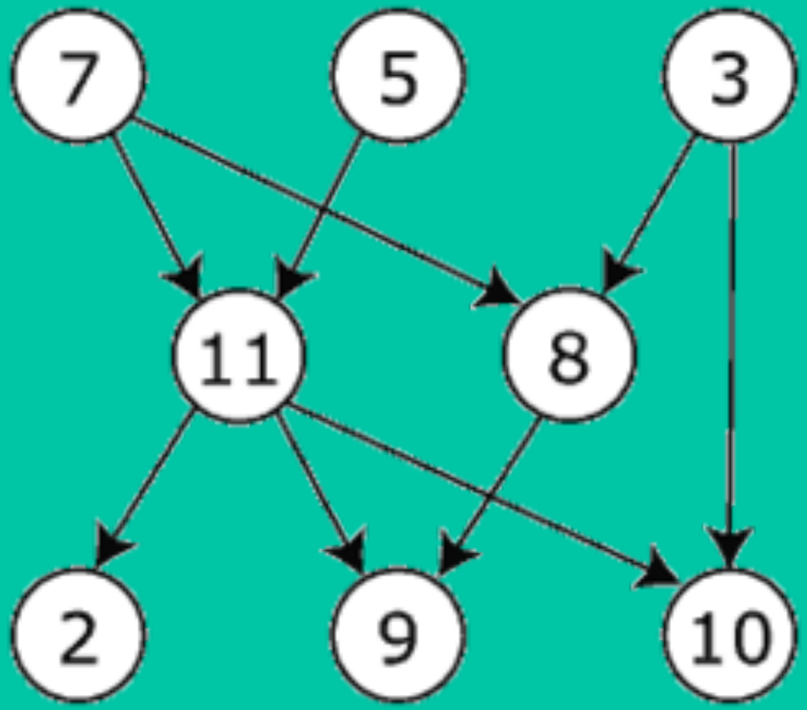


Красиво нарисованный процесс ETL из открытого источника (гугл поиск по картинкам)

tech

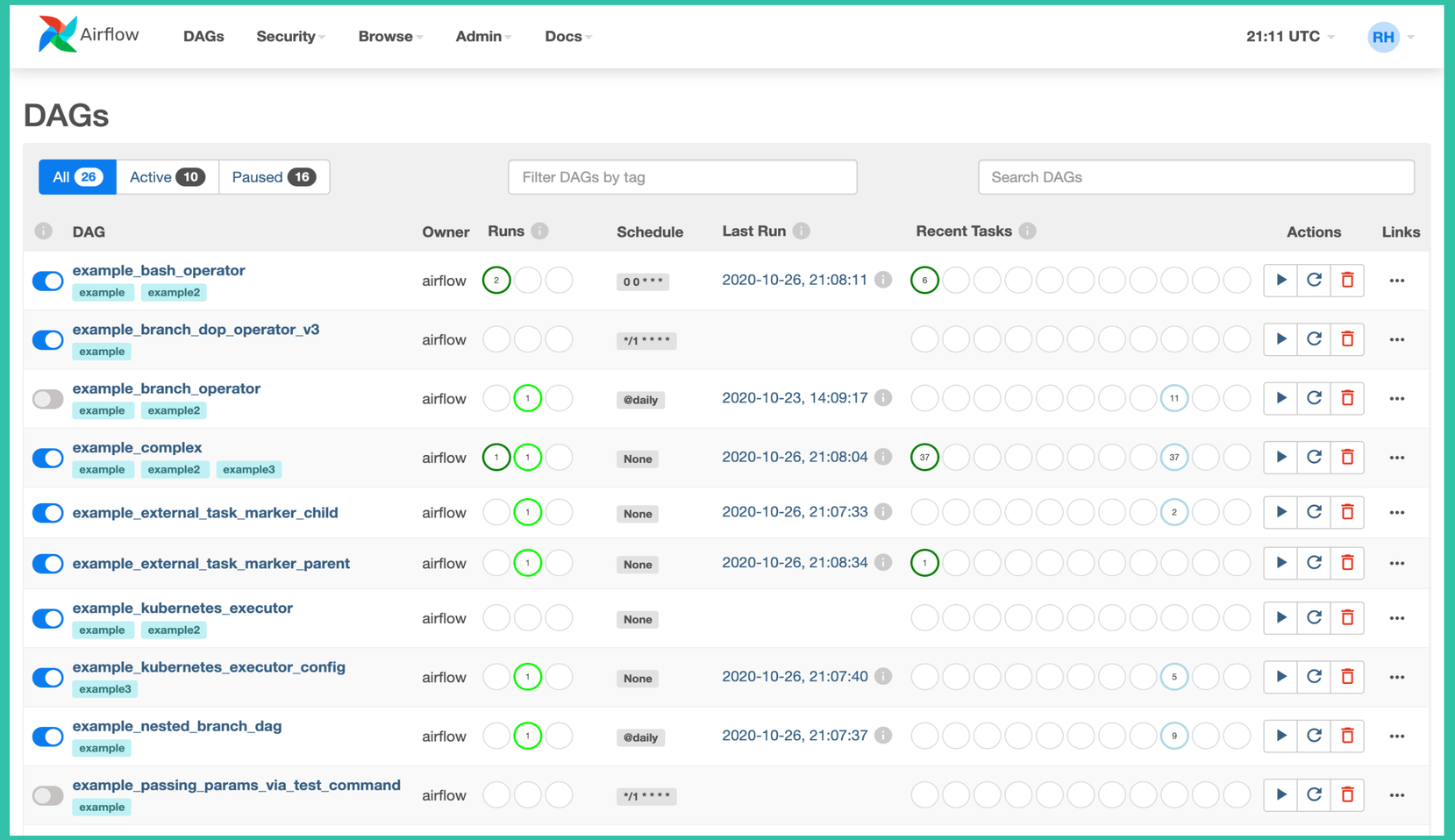
что такое DAG

Directed Acyclic Graph



Направленный ациклический граф (с) википедия

общий вид интерфейса airflow

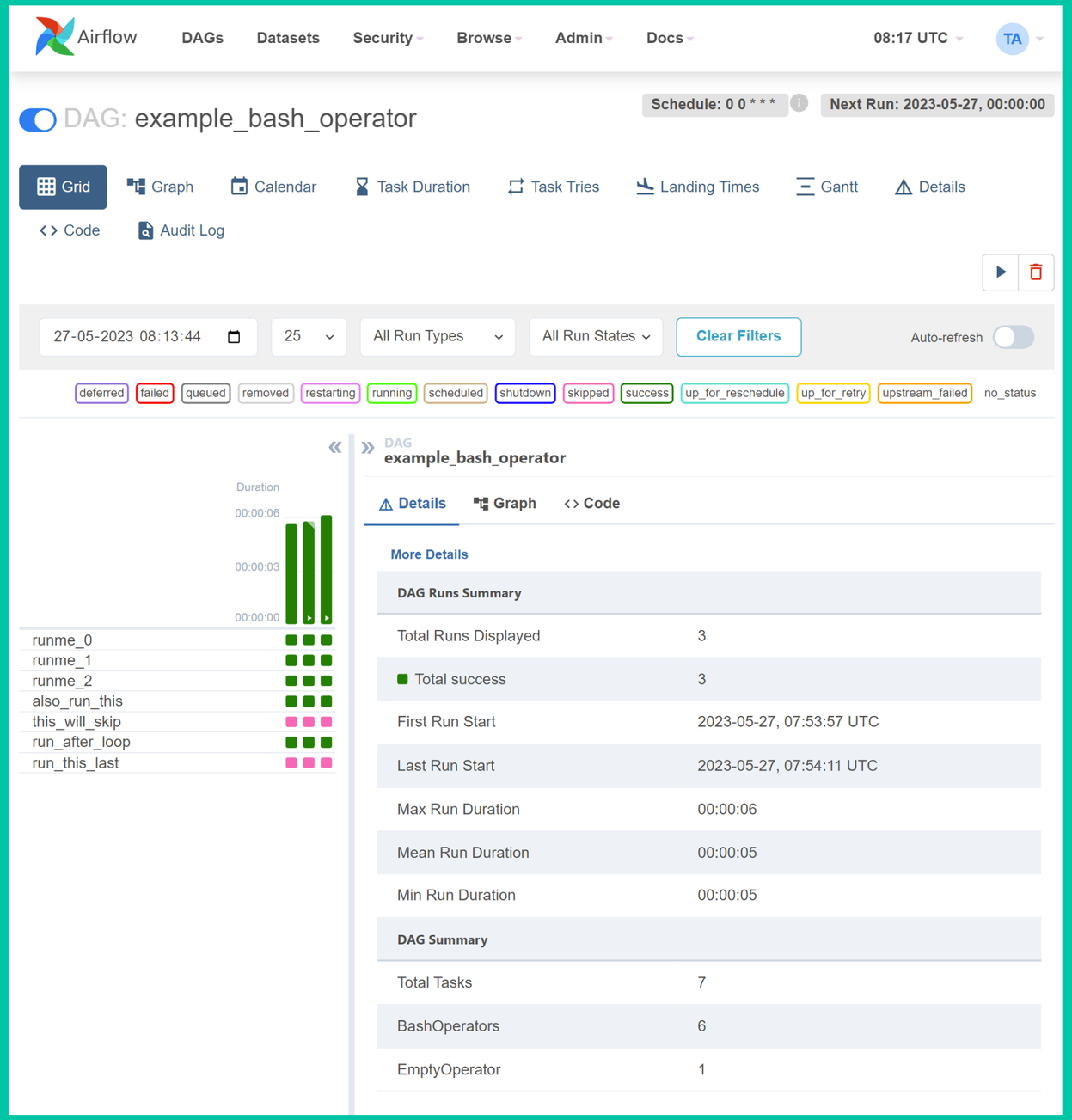


- список дагов
- описание догов
- статус тасок

tech

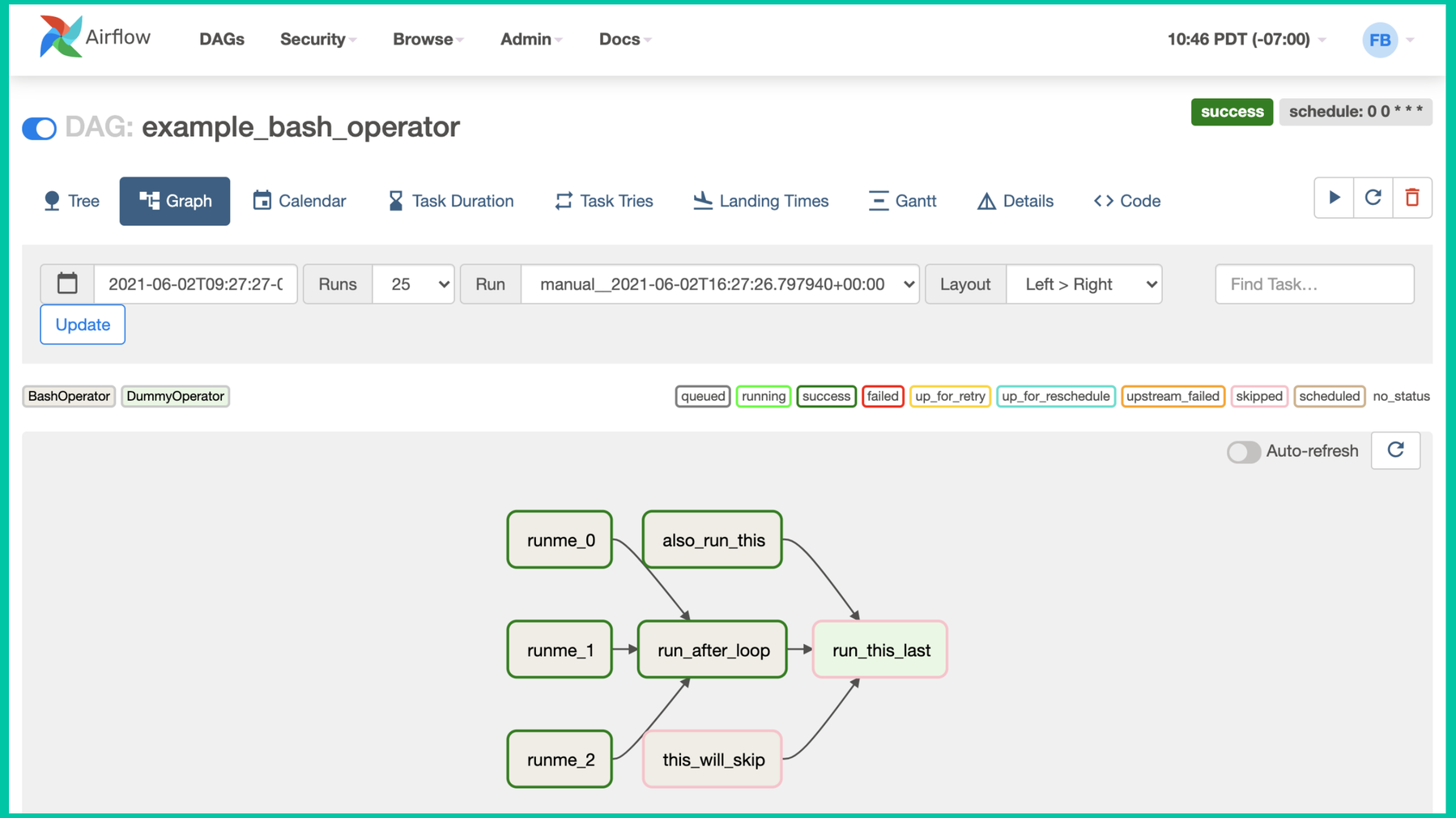
grid view

отображение по времени исполнения и статус конкретных тасок во всем даге



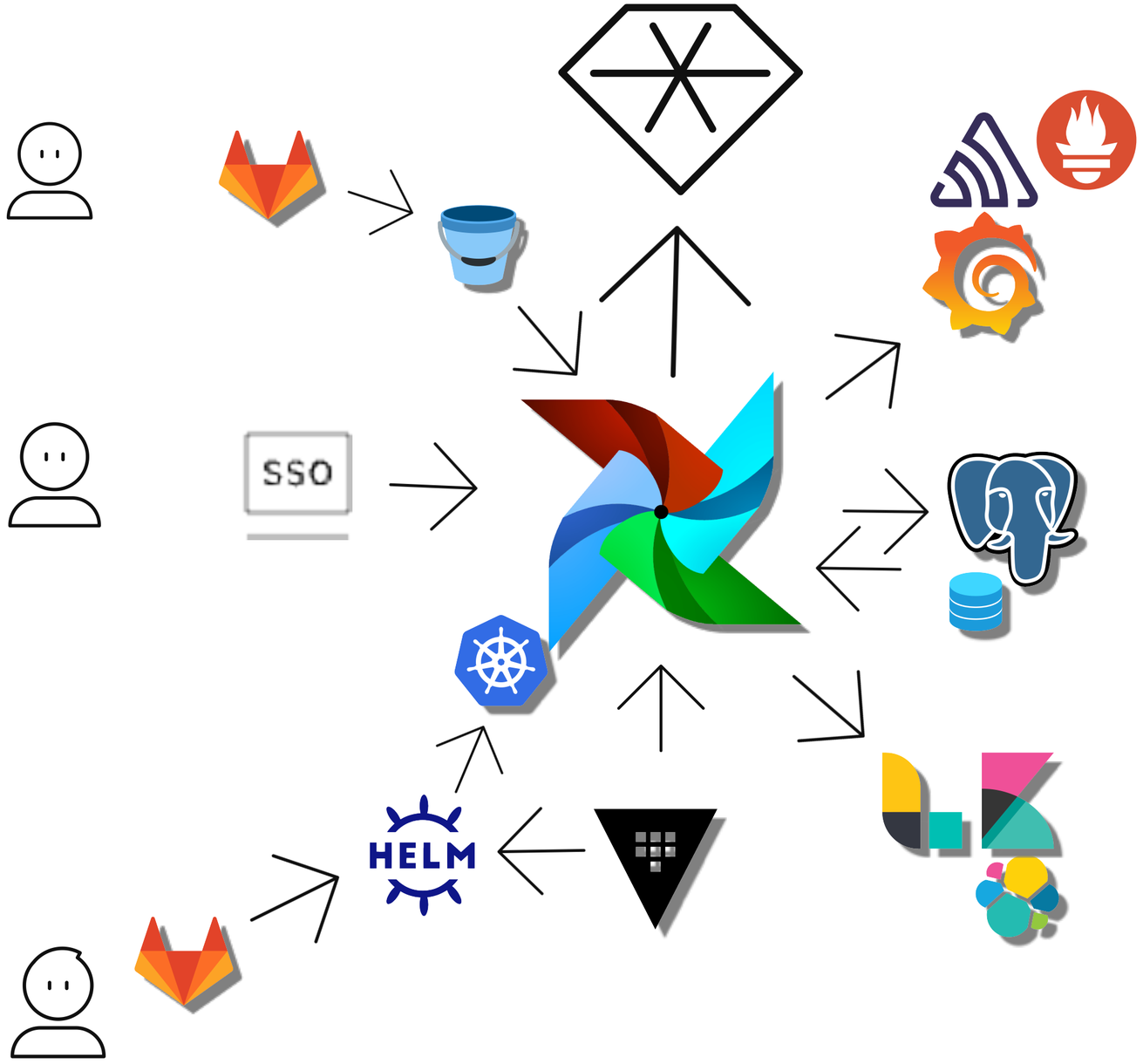
tech

graph view



отображение зависимостей тасок в даге

tech



Плюсы

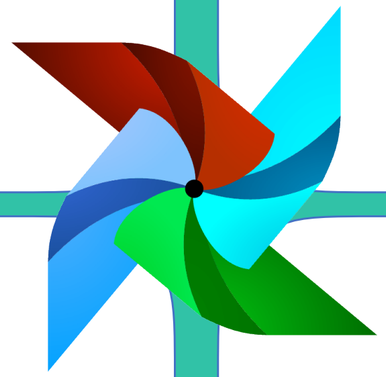
+ Airflow запускается предсказуемо и маштабируемо (kubernetes+helm), в стабильном окружении(gitlab).
+ Есть ревью дагов и тестовые среды перед продом.
+ Подключен secret backend для connections и секретов конфигурации(Hashicrop Vault).
+ Отдельно вынесенный кластер БД с метаданными(postgres).
+ Вход по SSO(ADFS).
+ Подключен log backend(elastic).
+ Выводятся метрики + алертинг (statsd+GrafanaOnCal).

Минусы

- Сложное добавление коннекшенов\переменных.
- Увеличение времени между созданием дага и запуском в прод.

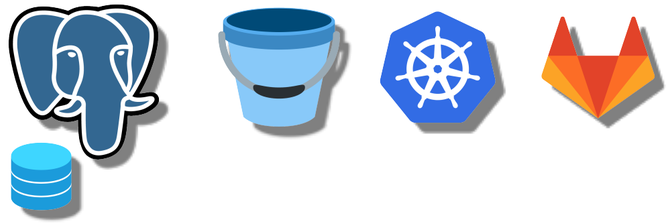
Observability

Security



VCS

High availability



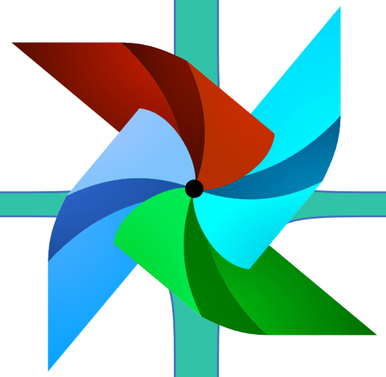
tech

Плюсы

Быстрое восстановление
Маштабируемость
Отказоустойчивость

Минусы

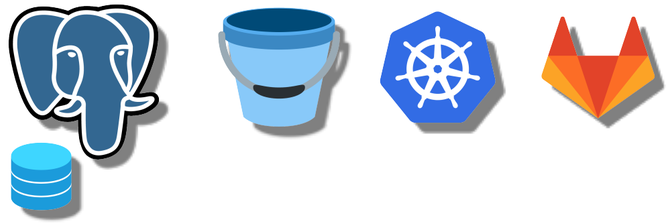
Увеличение времени доставки кода



Описание

Постгрес - хранение метаданных
s3 bucket - хранение логов и кода
k8s - инфраструктура для запуска приложения
gitlab - быстрое восстановления работоспособного состояния приложения и кода при сбоях.

High availability



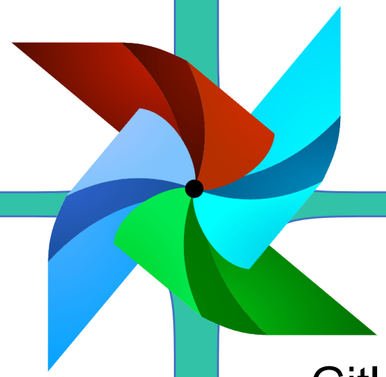
tech

Плюсы

Код DAGов и приложения версионируется, проверяется, аппрувается,
можно спокойно вести разработку без боязни положить прод, так как есть ci\cd и пачка сред

Минусы

Код медленнее попадает на прод, усложняется поддержка приложения



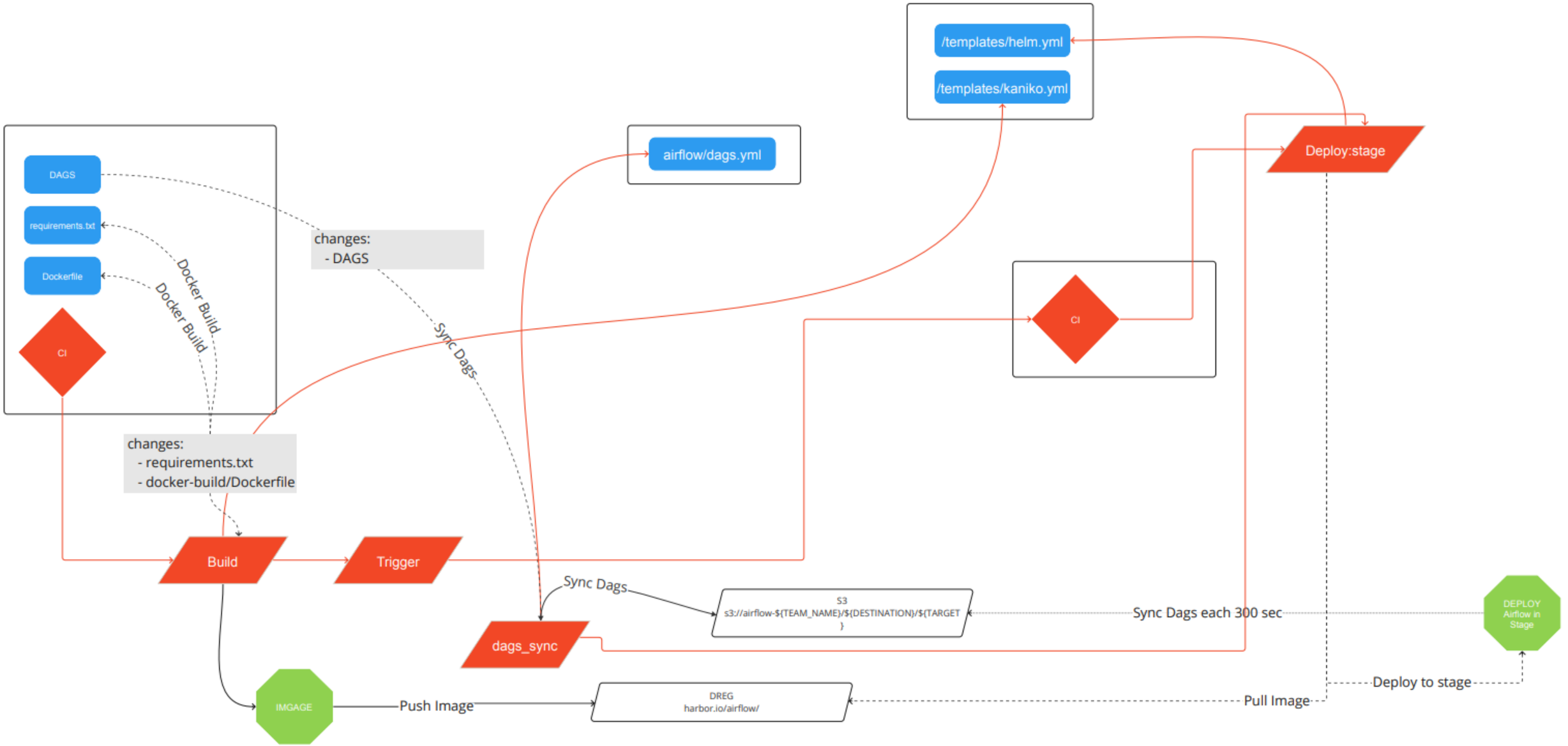
VCS

Описание

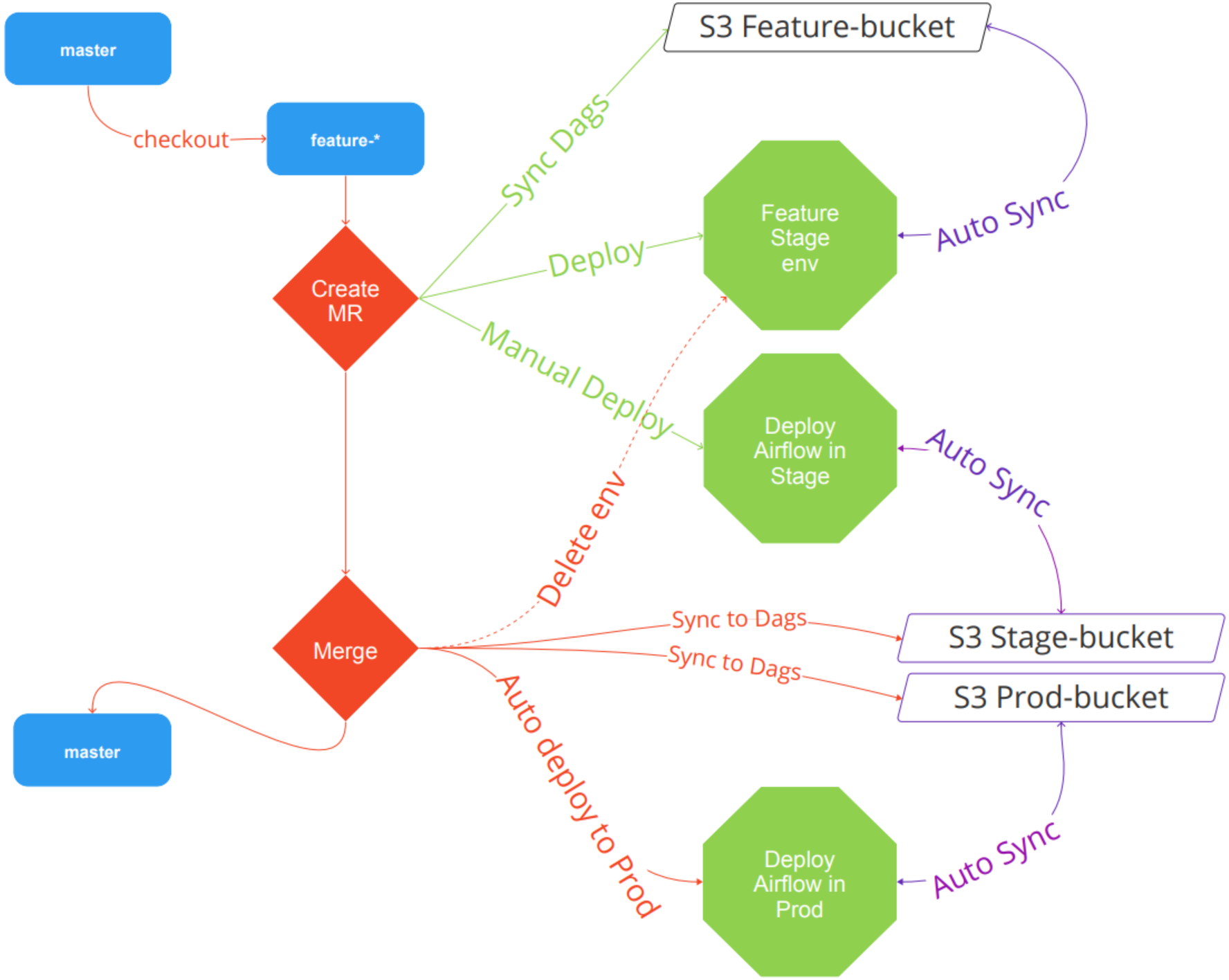
Gitlab - ci\cd+git
Helm - темплейтирование приложения

tech

На примере Stage ветки



changes:
- requirements.txt
- docker-build/Dockerfile

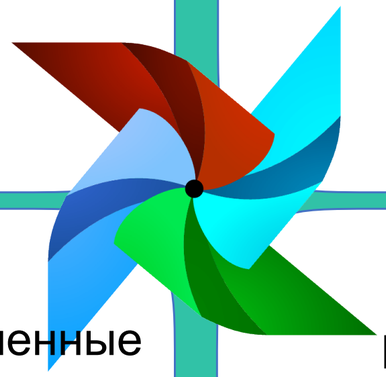


tech

Observability

Описание

StatsD - экспорт метрик
Elastic Stack - сбор и хранение логов + поиск
Sentry - сбор и визуализация трейсов
Grafana - визуализация метрик



Плюсы

Айрфлоу отдает метрики и логи в предназначенные для этого системы, взамен мы получаем наблюдаемость и вдумчивый дебаг

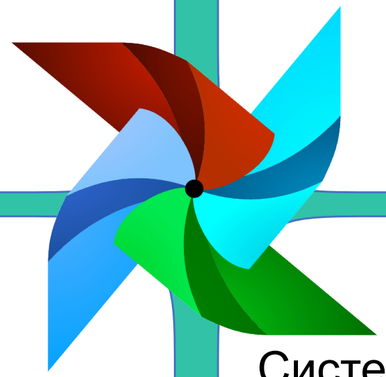
Минусы

Поддержка этих сервисов требует внимания.

Описание

Hashicorp Vault - хранение и создание секретов
SSO - доступ по LDAP
Gitlab - хранение ci переменных, ограничение доступа к ci и репозитория по группам

Security



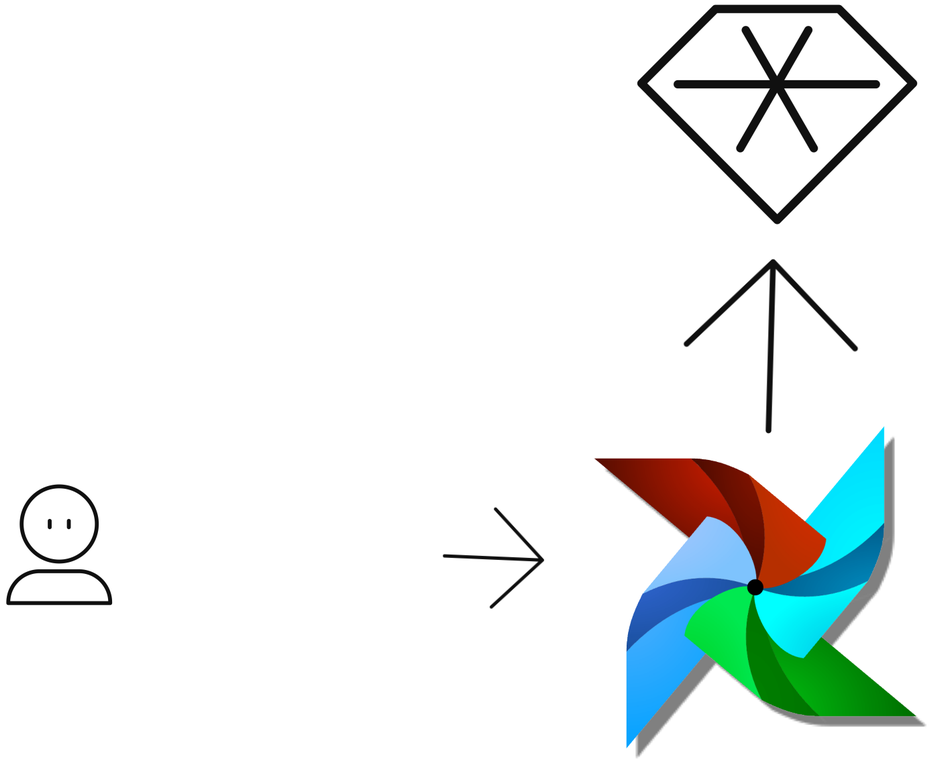
Плюсы

Секреты версионируются и отслеживаются, появляется возможность автоматического обновления секретов.
Убирается ручное управление пользователями и их группами

Минусы

Системы требуют поддержки.
Медленное изменение коннекшенов по сравнению с UI airflow

tech

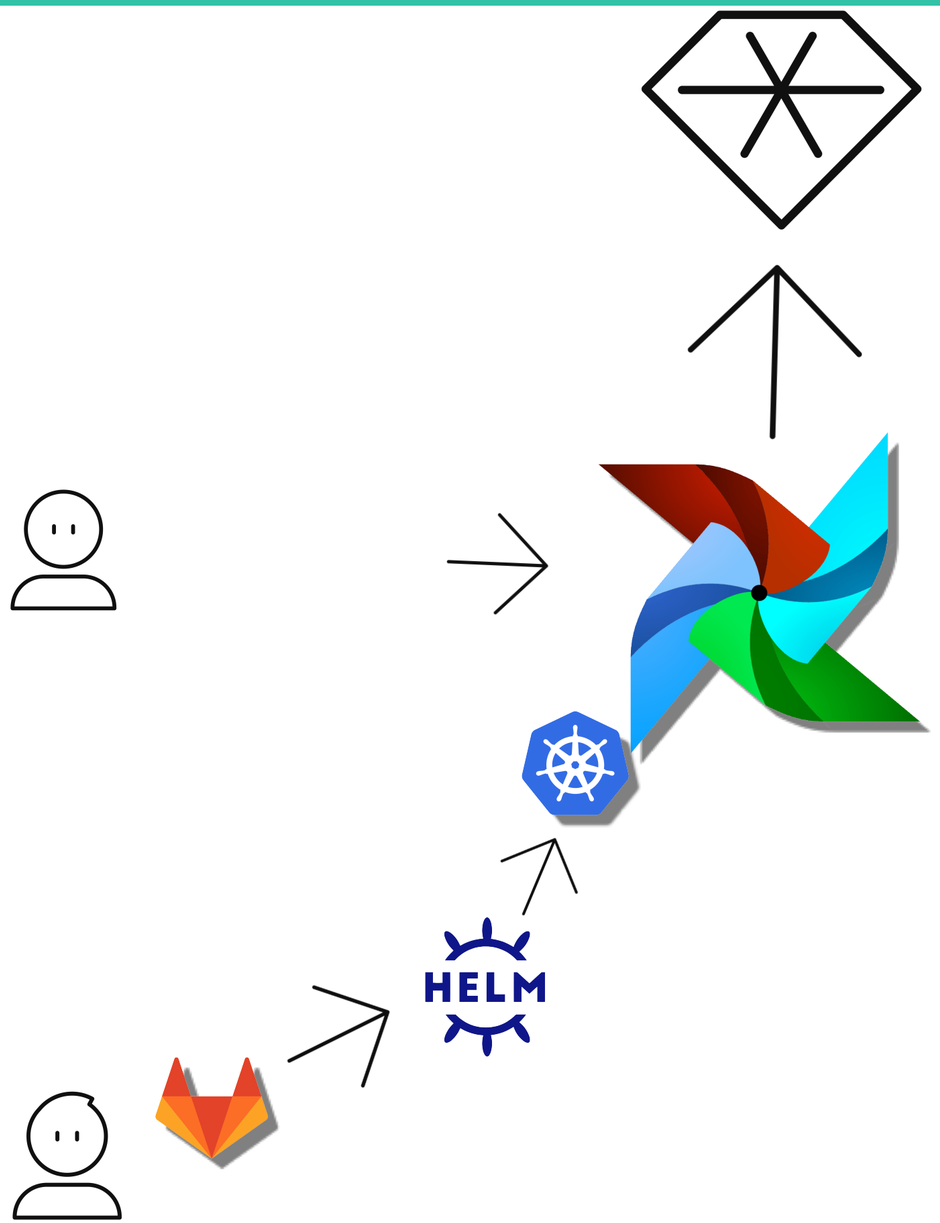


Плюсы

+ Airflow работает, ETL процесс запущены
+ Быстрая выкатка дагов на прод
+ Легкое добавление секретов и переменных

Минусы

- Пользователи запускают Airflow локально
- Окружение отличается от локалки к локалке
- Немаштабируемо
- Даги не ревьювятся, хранятся в “надежном” месте
- Секреты хранятся в “надежном” месте и их знает немного ограниченный круг лиц
- Логи локальные
- Метрики локальные
- БД с метаданными локальная

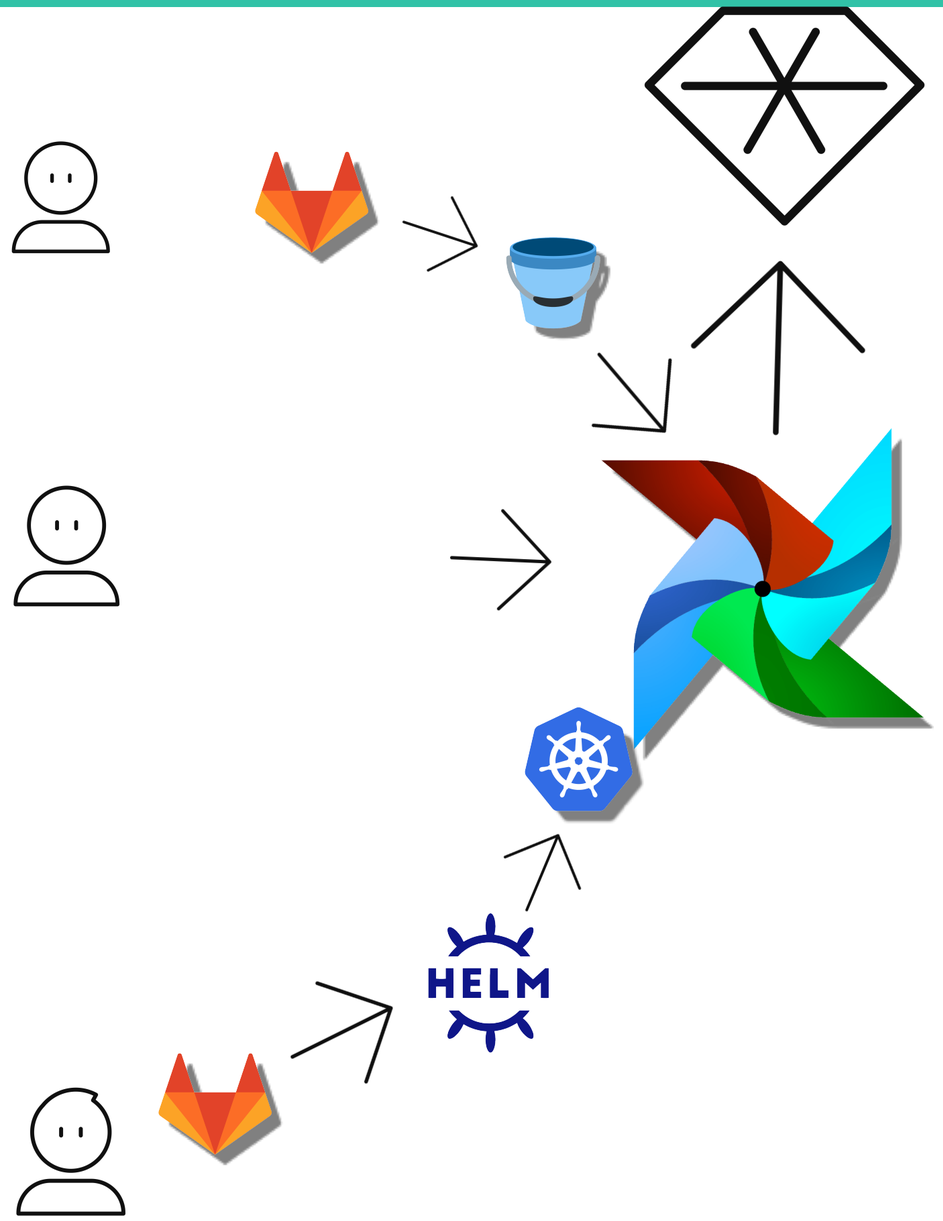


Плюсы

+ Airflow запускается предсказуемо и маштабируемо(kubernetes+helm)
+ Быстрая выкатка дагов на прод
+ Легкое добавление секретов и переменных

Минусы

- Даги не ревьювятся, хранятся в “надежном” месте
- Секреты хранятся в “надежном” месте
- Логи локальные
- Метрики локальные.
- БД с метаданными локальная



Плюсы

+ Airflow запускается предсказуемо и маштабируемо(kubernetes+helm), в стабильном окружении(gitlab).

+ Есть ревью дагов и тестовые среды перед продом

+ Легкое добавление секретов и переменных

Минусы

– Увеличение времени между созданием дага и запуском в прод

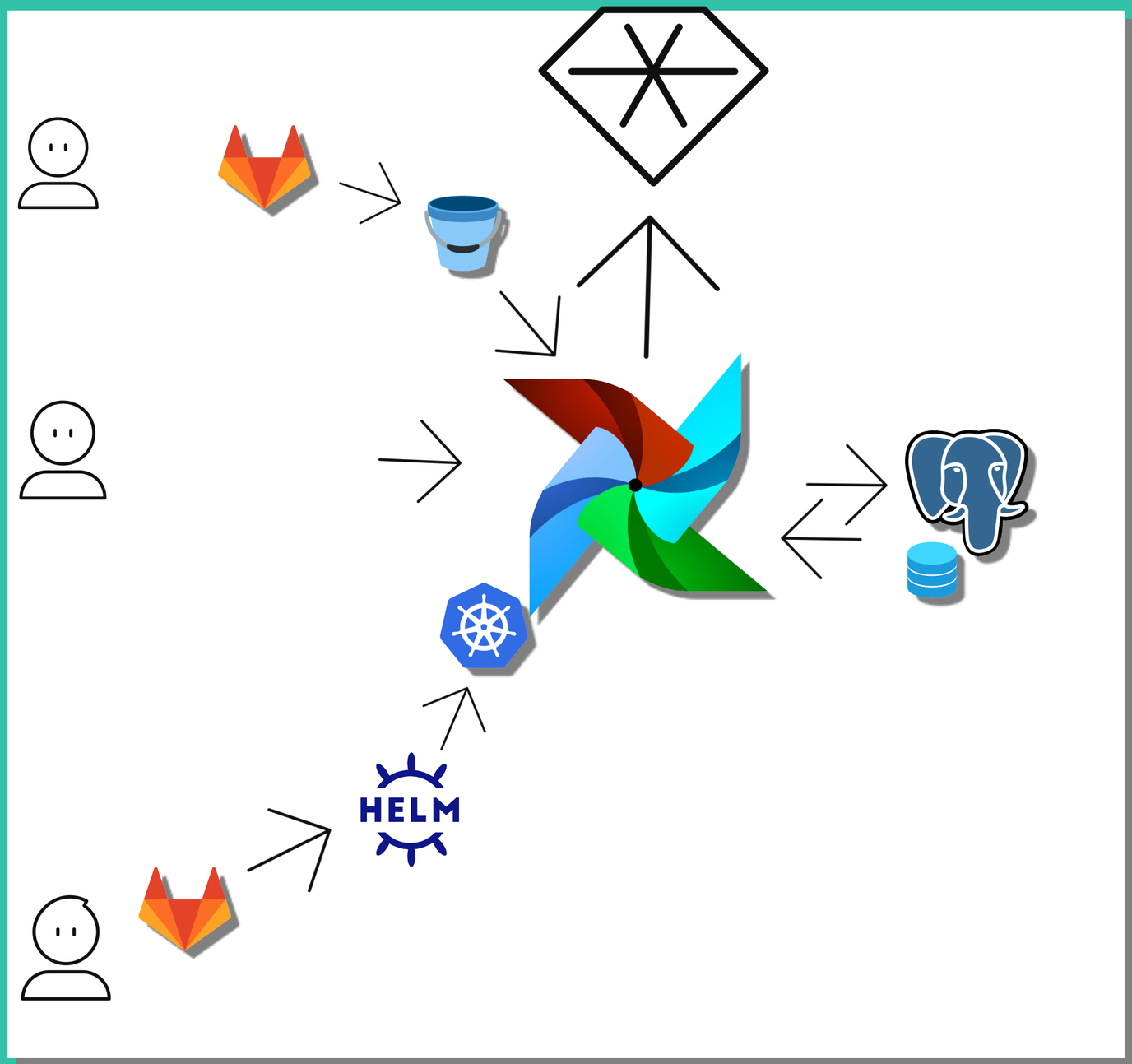
– Ненадежно

– Секреты хранятся в “надежном” месте

– Логи локальные

– Метрики локальные.

– БД с метаданными локальная



Плюсы

+ Airflow запускается предсказуемо и маштабируемо(kubernetes+helm), в стабильном окружении(gitlab).

+ Есть ревью дагов и тестовые среды перед продом

+ Легкое добавление секретов и переменных

+ Отдельно вынесенный кластер БД с метаданными(postgres)

Минусы

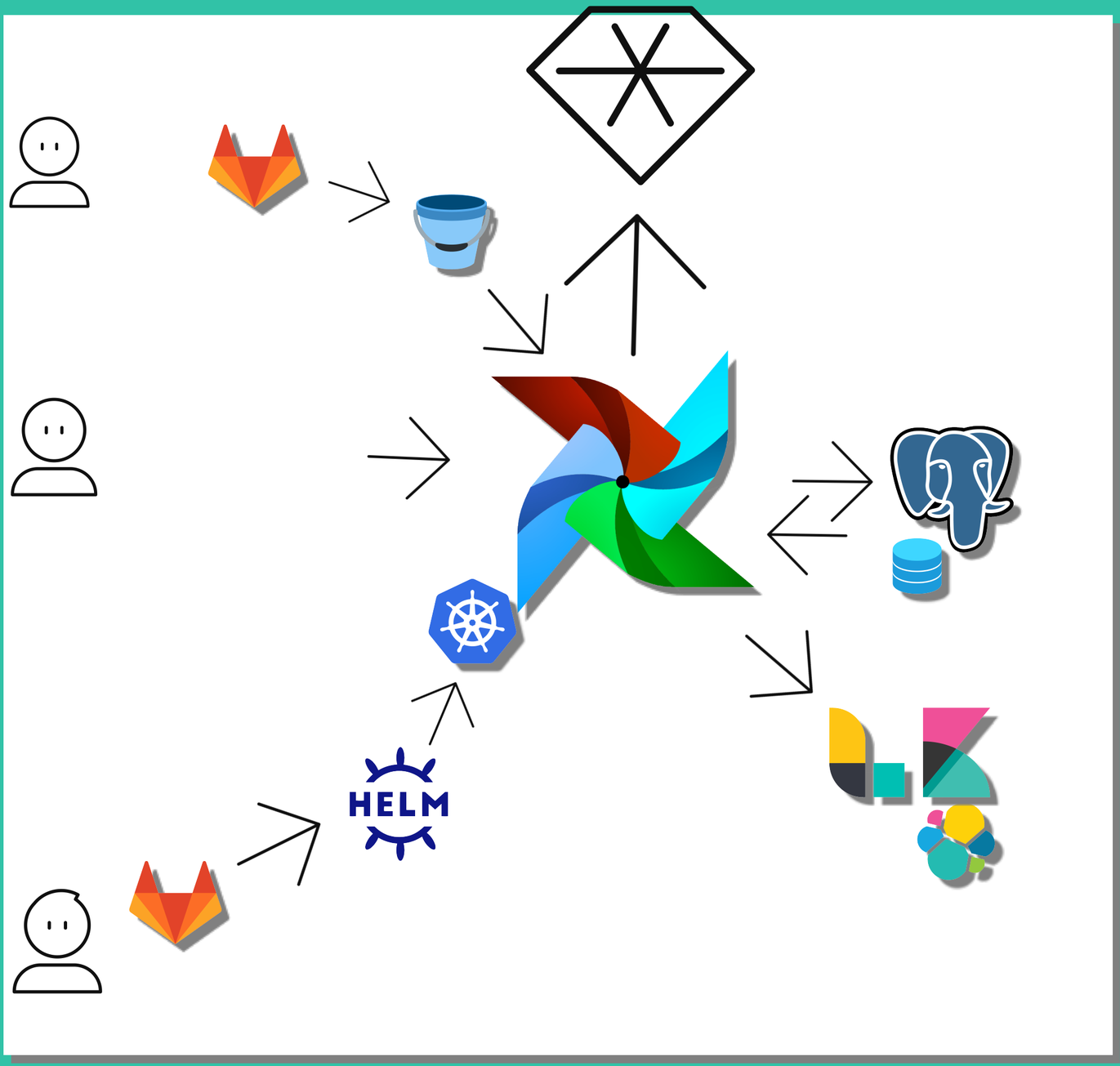
– Увеличение времени между созданием дага и запуском в прод

– Ненадежно

– Секреты хранятся в “надежном” месте

– Логи локальные

– Метрики локальные



Плюсы

+ Airflow запускается предсказуемо и маштабируемо(kubernetes+helm), в стабильном окружении(gitlab).

+ Есть ревью дагов и тестовые среды перед продом

+ Легкое добавление секретов и переменных

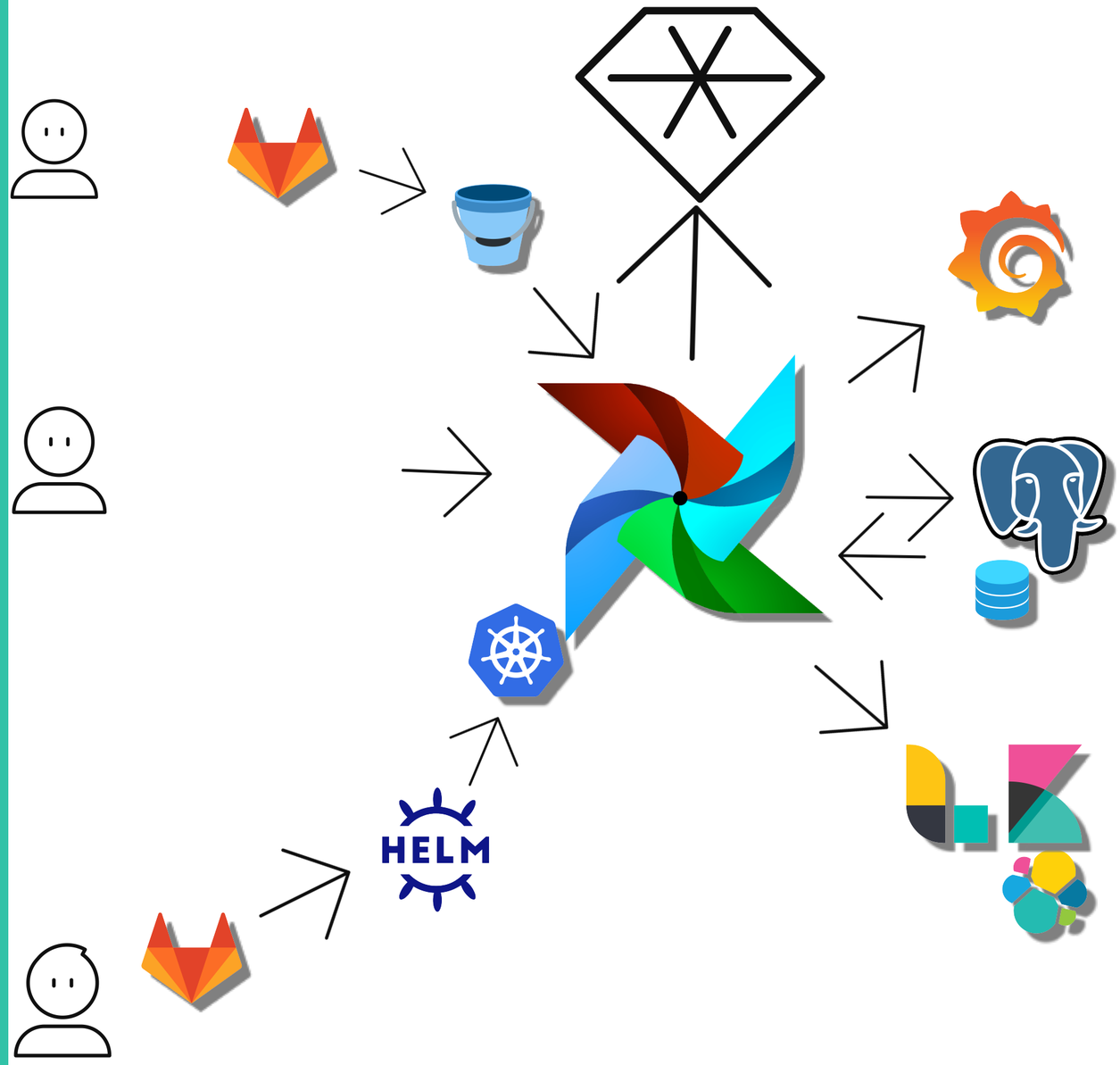
+ Отдельно вынесенный кластер БД с метаданными(postgres)

+ Подключен log backend(elastic)

Минусы

– Увеличение времени между созданием дага и запуском в прод

– Секреты хранятся в надежном месте



Плюсы

+ Airflow запускается предсказуемо и маштабируемо(kubernetes+helm), в стабильном окружении(gitlab).

+ Есть ревью дагов и тестовые среды перед продом

+ Легкое добавление секретов и переменных

+ Отдельно вынесенный кластер БД с метаданными(postgres)

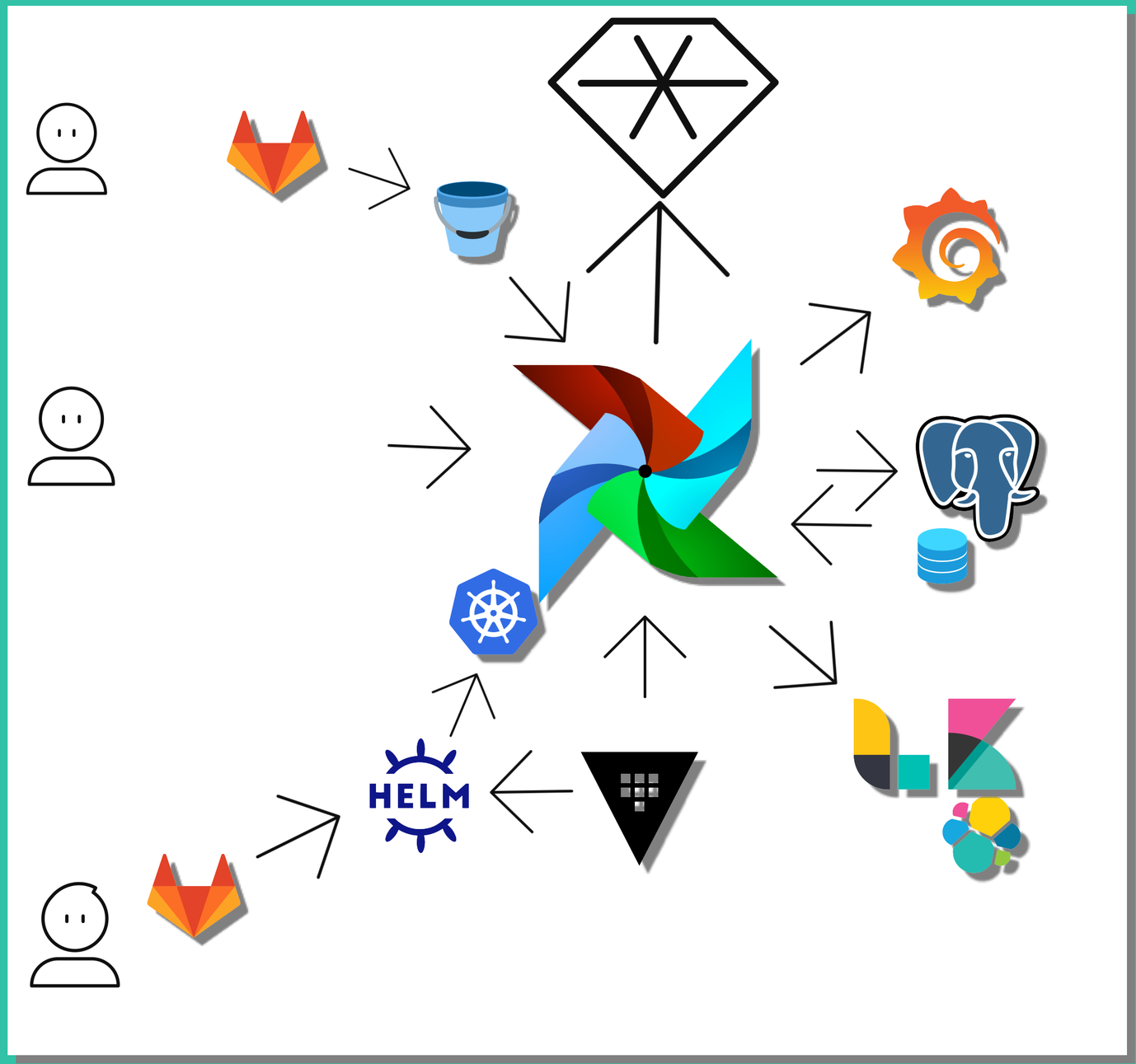
+ Подключен log backend(elastic)

+ Выводятся метрики + алертинг (statsd+GrafanaOnCal)

Минусы

− Увеличение времени между созданием дага и запуском в прод

− Секреты хранятся в надежном месте



Плюсы

+ Airflow запускается предсказуемо и маштабируемо(kubernetes+helm), в стабильном окружении(gitlab).

+ Есть ревью дагов и тестовые среды перед продом

+ Подключен secret backend для connections и секретов конфигурации(Hashicrop Vault).

+ Отдельно вынесенный кластер БД с метаданными(postgres)

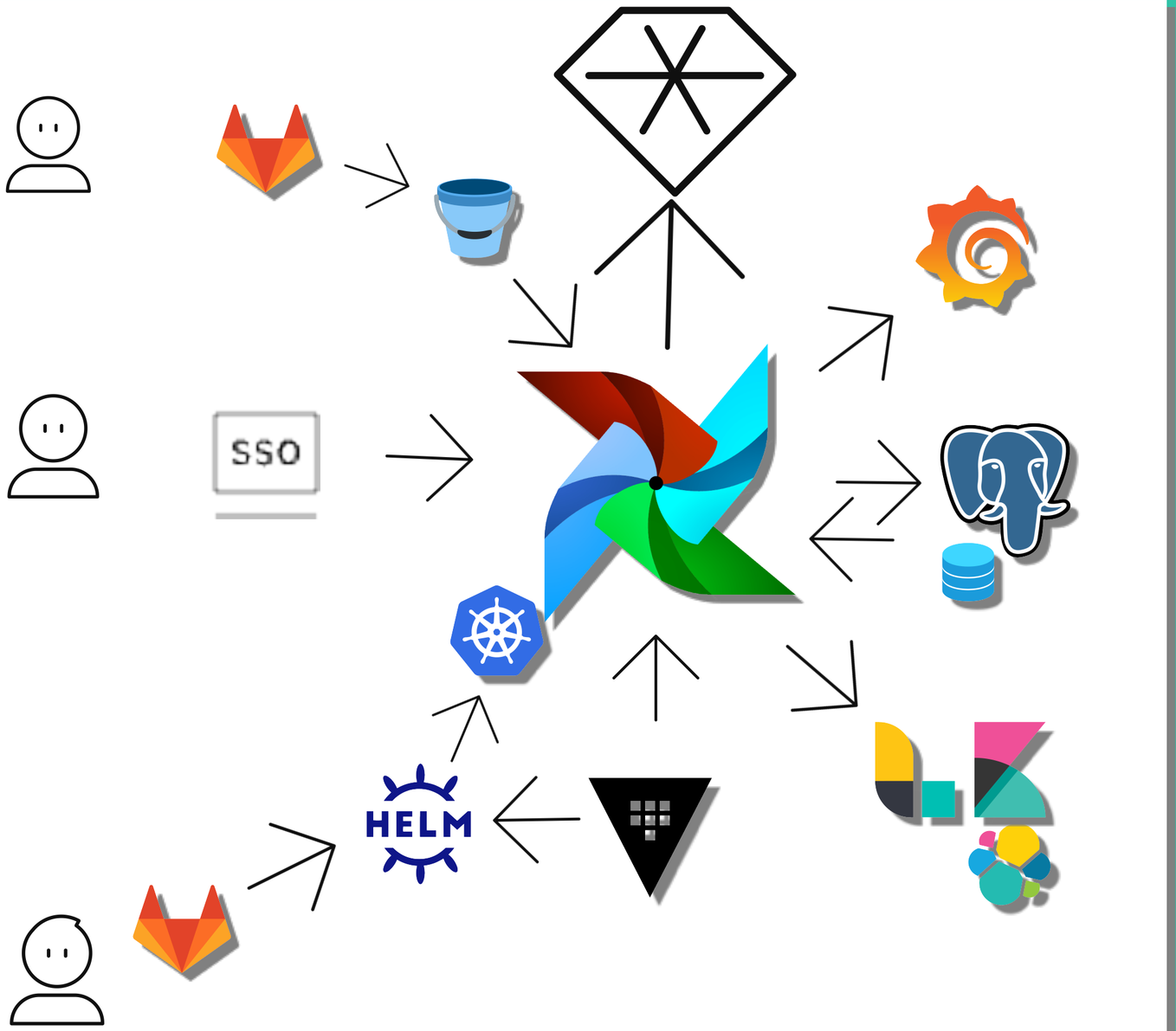
+ Подключен log backend(elastic)

+ Выводятся метрики + алертинг (statsd+GrafanaOnCal)

Минусы

– Сложное добавление коннекшенов\переменных.

– Увеличение времени между созданием дага и запуском в прод



Плюсы

+ Airflow запускается предсказуемо и маштабируемо(kubernetes+helm), в стабильном окружении(gitlab).

+ Есть ревью дагов и тестовые среды перед продом

+ Подключен secret backend для connections и секретов конфигурации(Hashicrop Vault).

+ Отдельно вынесенный кластер БД с метаданными(postgres)

+ Вход по SSO(ADFS)

+ Подключен log backend(elastic)

+ Выводятся метрики + алертинг (statsd+GrafanaOnCal)

Минусы

− Сложное добавление коннекшенов\переменных.

− Увеличение времени между созданием дага и запуском в прод

ВЫВОДЫ

tech

спасибо за внимание

tech