

Airflow

Фёдор Ерин

Data Scientist

Ex BCG Gamma, X5 Retail Group, Mail.ru Group (VK)

Skillbox



Фёдор Ерин

Data Scientist

5 лет опыта

в анализе данных
и машинном обучении

3 года преподавания

онлайн-курсов по анализу
данных и дата инжинирингу

Образование —

МФТИ (Физтех),
ШАД (Яндекс)

Выполненные проекты

в ритейле, банкинге,
промышленности,
онлайн-продуктах

Цели модуля

1

Изучить основные
понятия и сущности Airflow

2

Научиться настраивать
и запускать графы
вычислений (DAG)

3

Разобрать настройку и отладку
пайплайна обработки данных

4

Освоить локальный запуск
Airflow и работу с ним

Airflow

Введение в Airflow

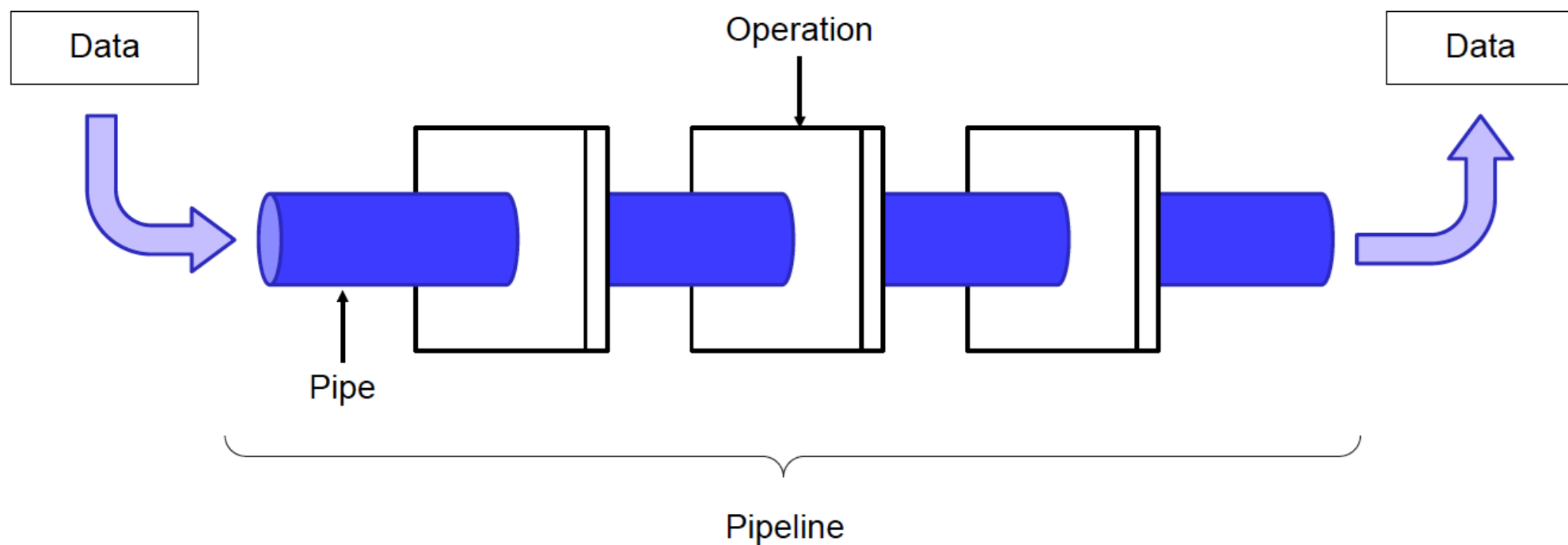
Skillbox

образовательная платформа

Цель видео

Изучить основные понятия и сущности, разобрать архитектуру Airflow.

Конвейер задач



Что такое Airflow?

Airflow — инструмент создания, мониторинга и оркестрации процессов обработки данных

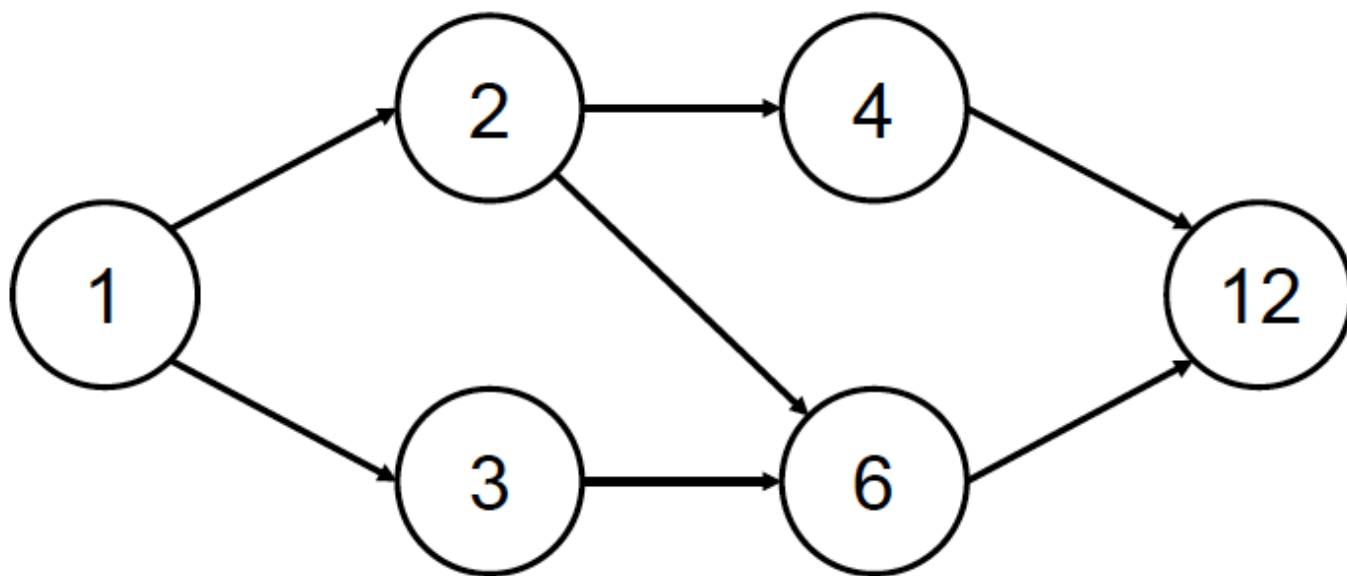
- Открытый исходный код
- Веб-интерфейс
- Написан на Python
- Широкое использование
- Удобен в применении
- Легко масштабируем
- Собственный репозиторий метаданных
- Интеграция со множеством источников и сервисов
- Расширяемый REST API



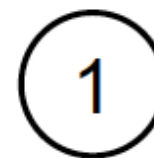
Представим пайплайн обработки данных

- 1 Выгрузить данные о продажах магазина из базы
- 2 Рассчитать выручку
- 3 Сохранить результат в файл

DAG (Directed Acyclic Graph) — направленный ациклический граф



Вершина



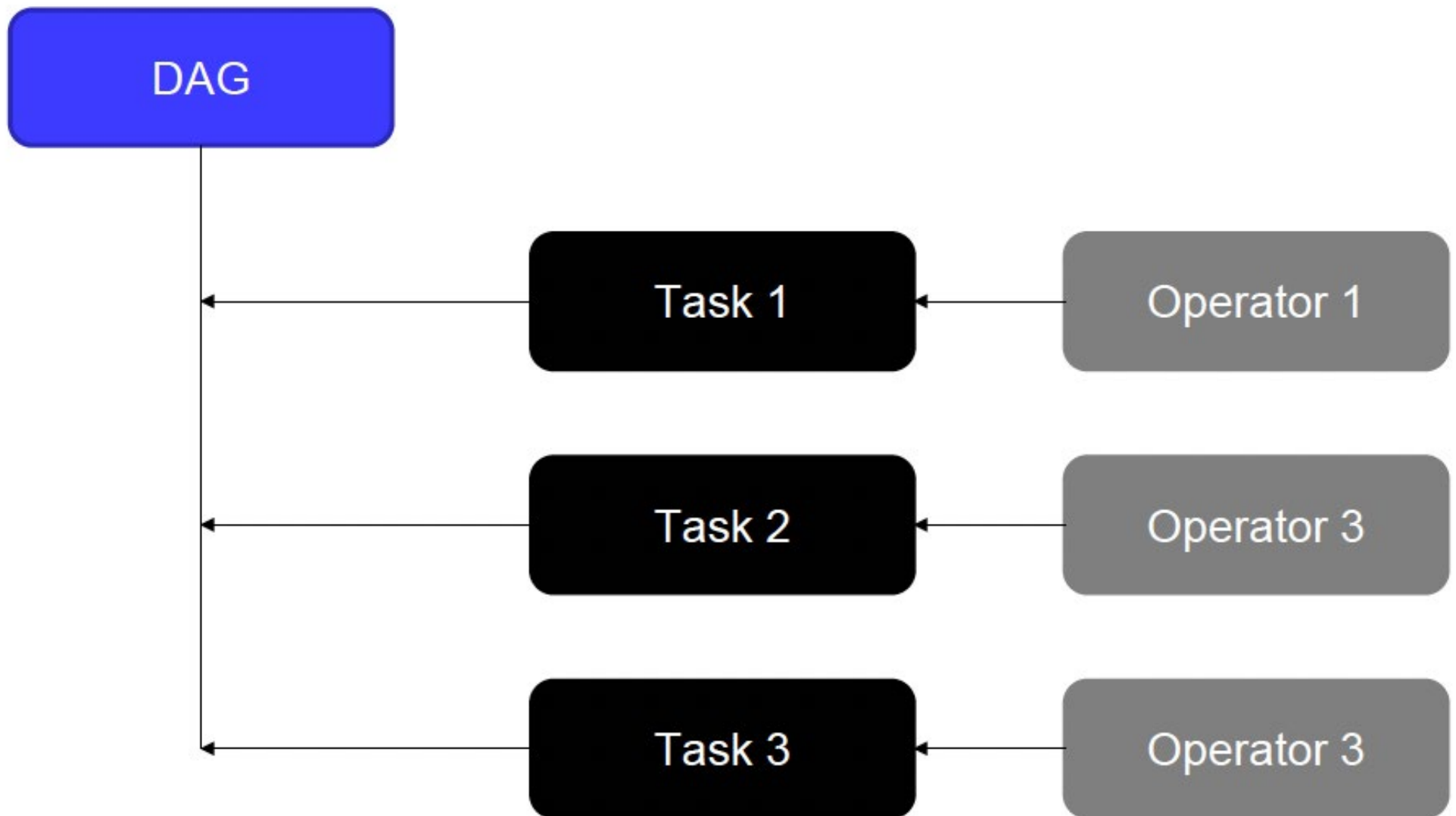
Направленная СВЯЗЬ



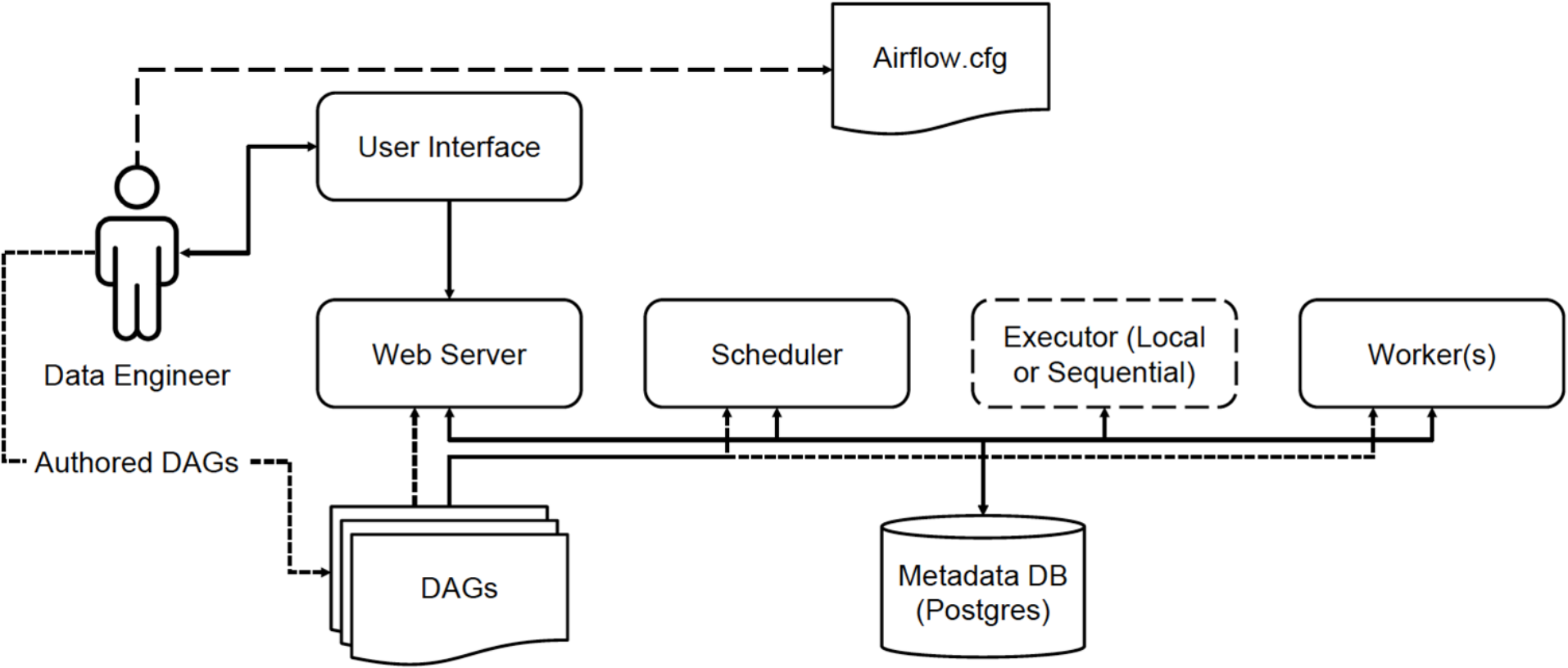
Ненаправленная СВЯЗЬ



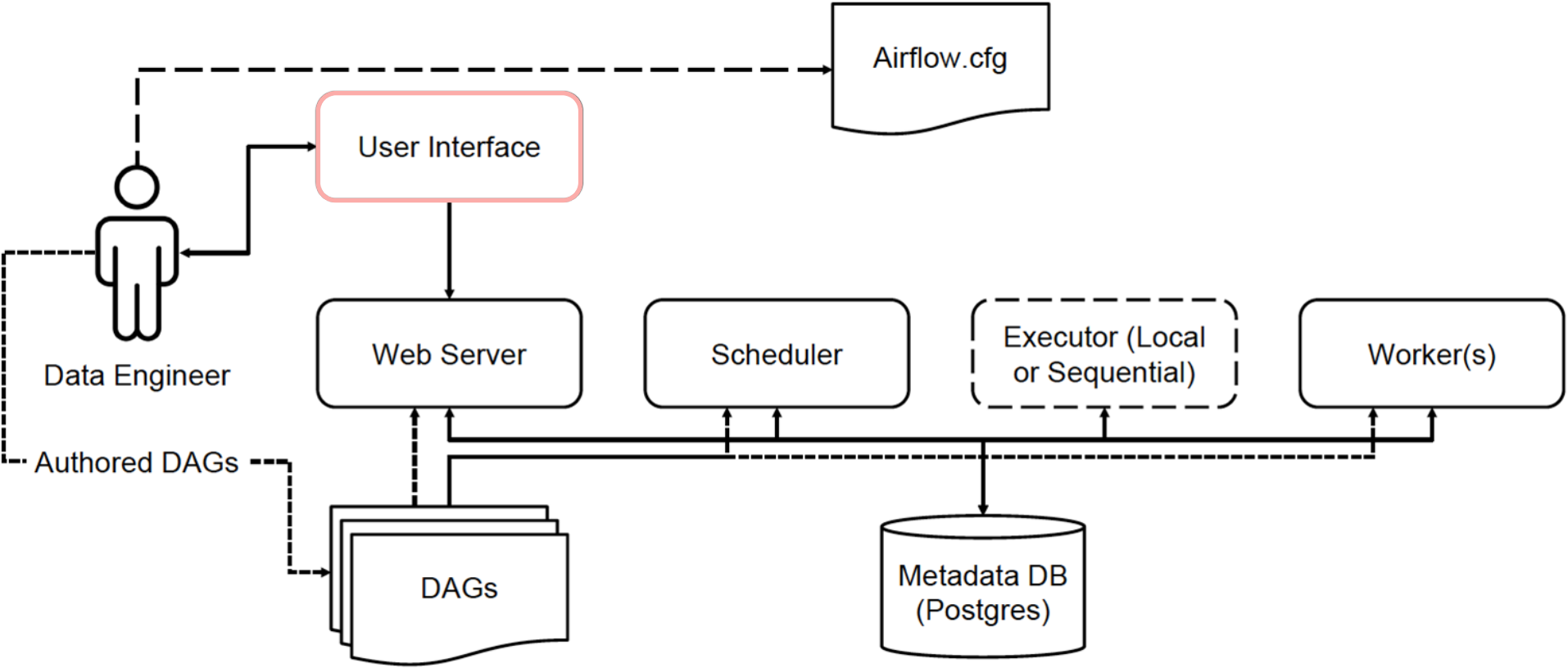
Operators, Tasks



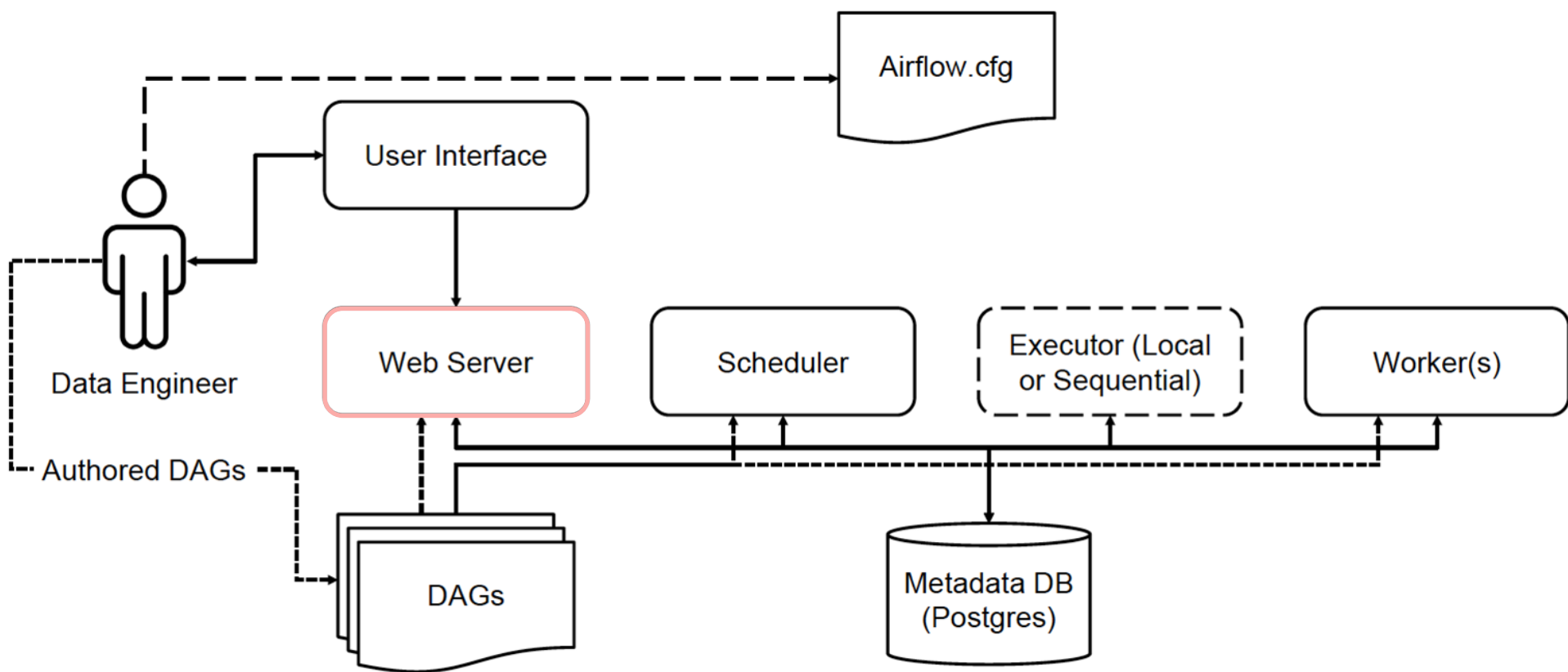
Архитектура



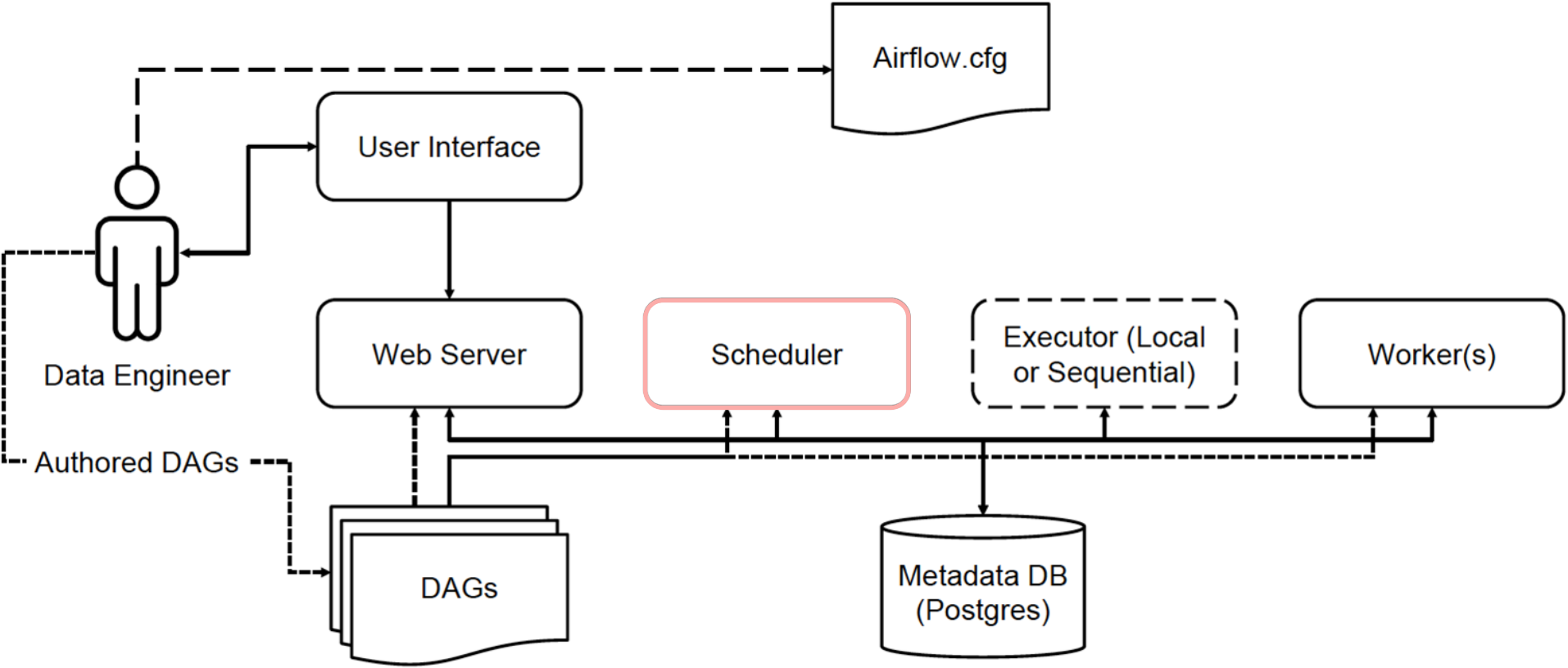
Архитектура



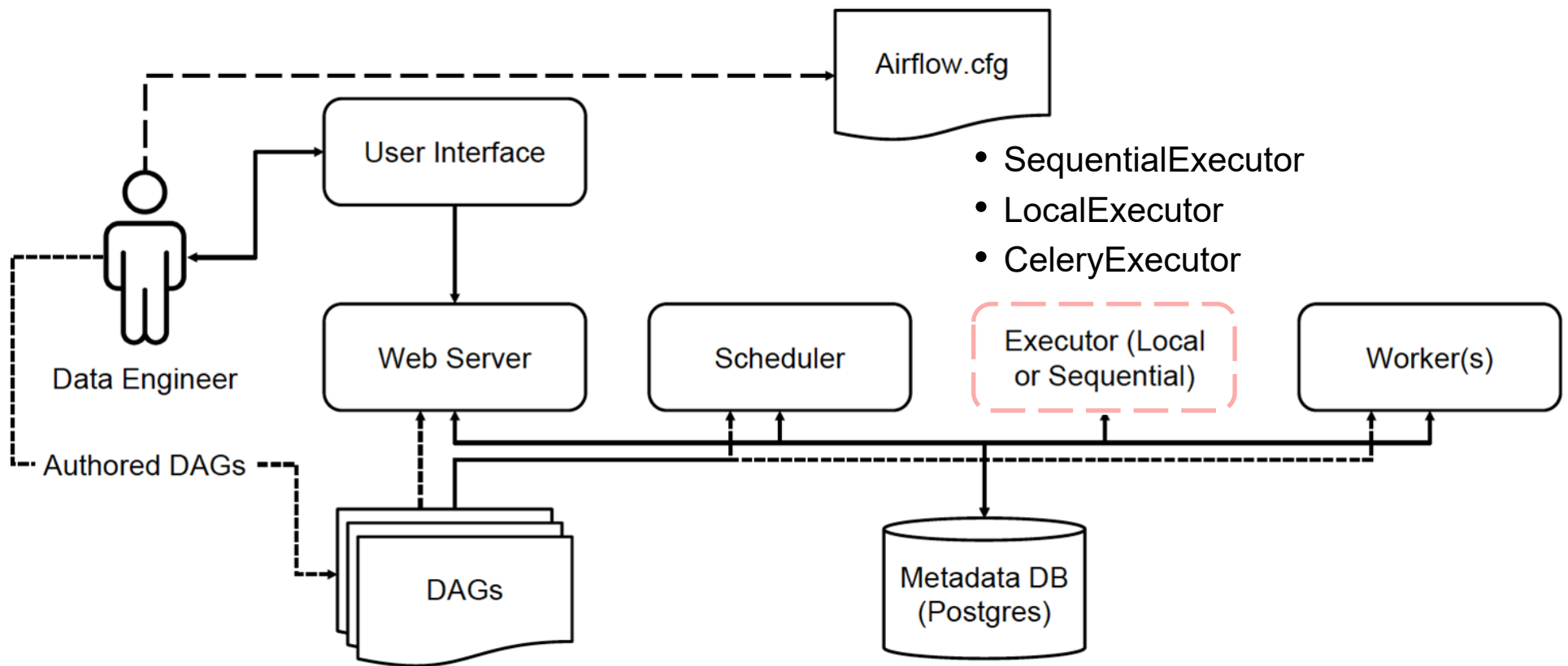
Архитектура



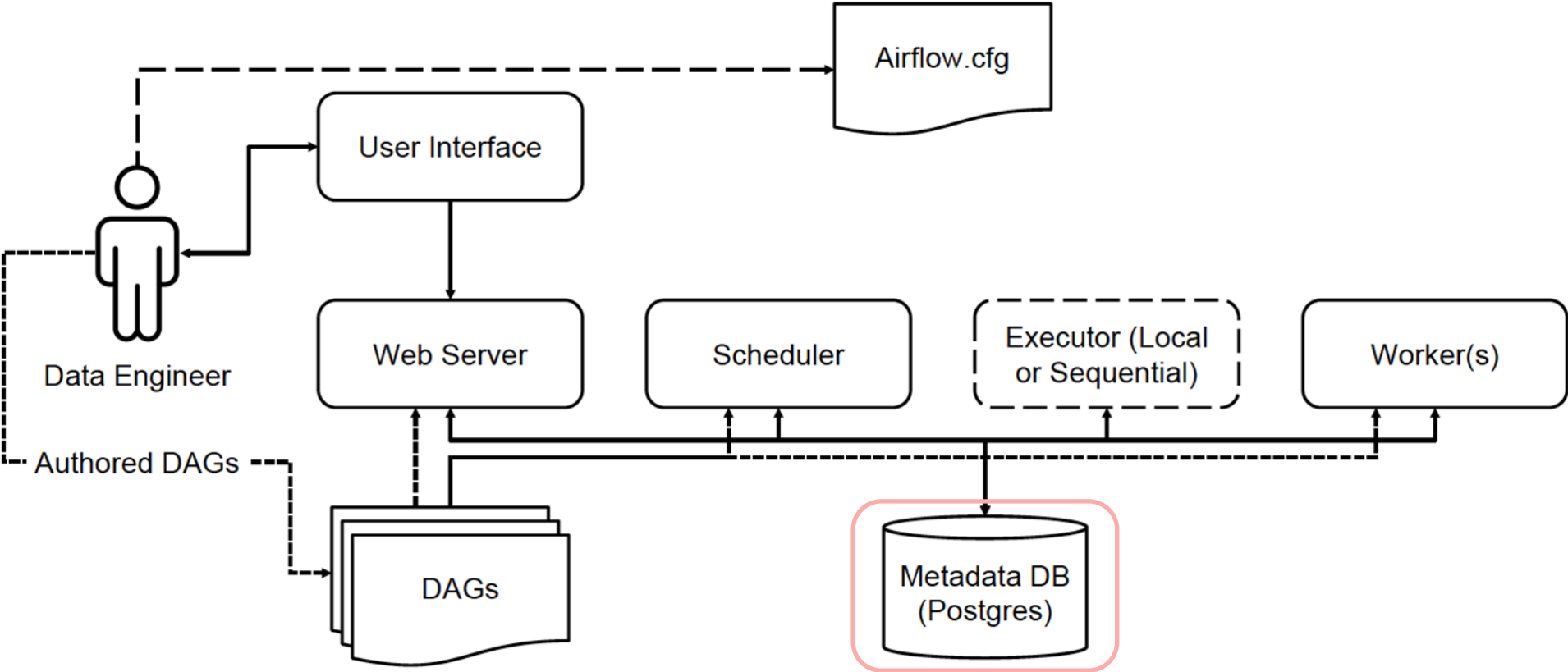
Архитектура



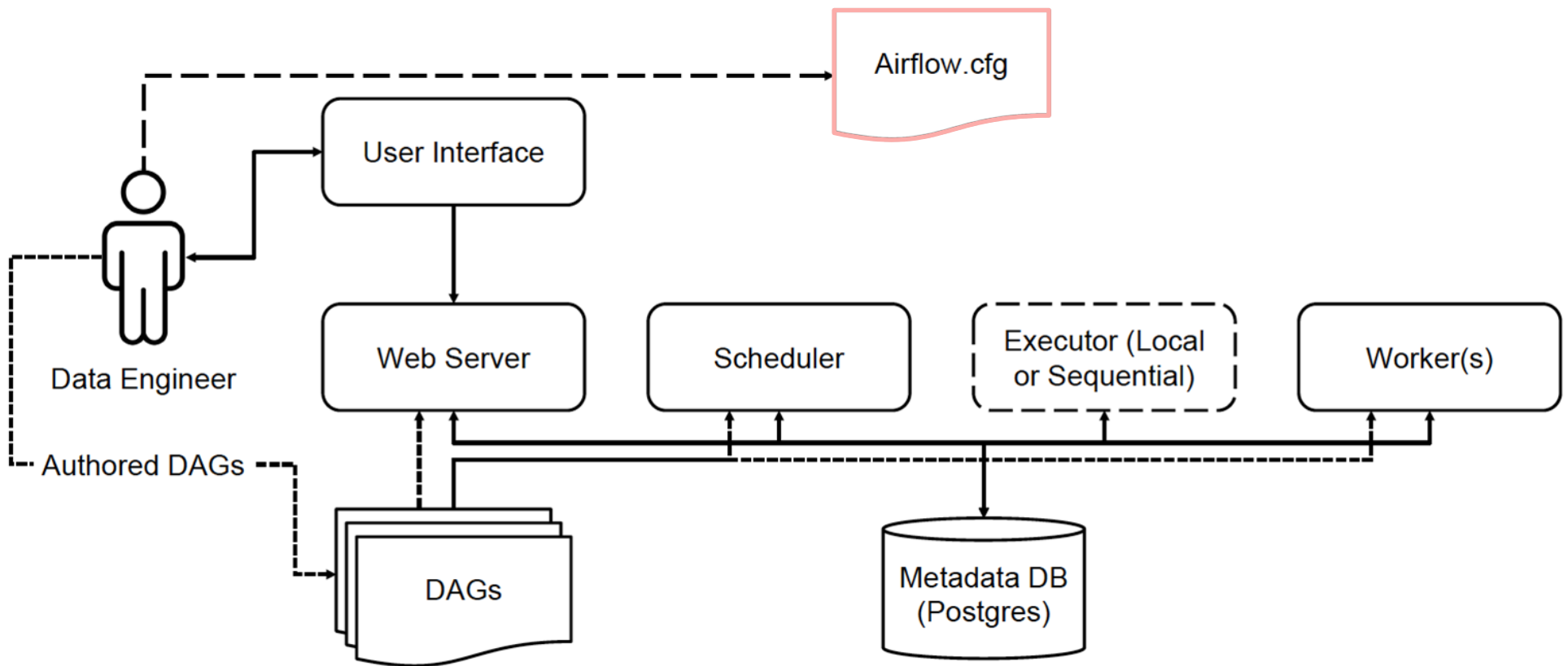
Архитектура



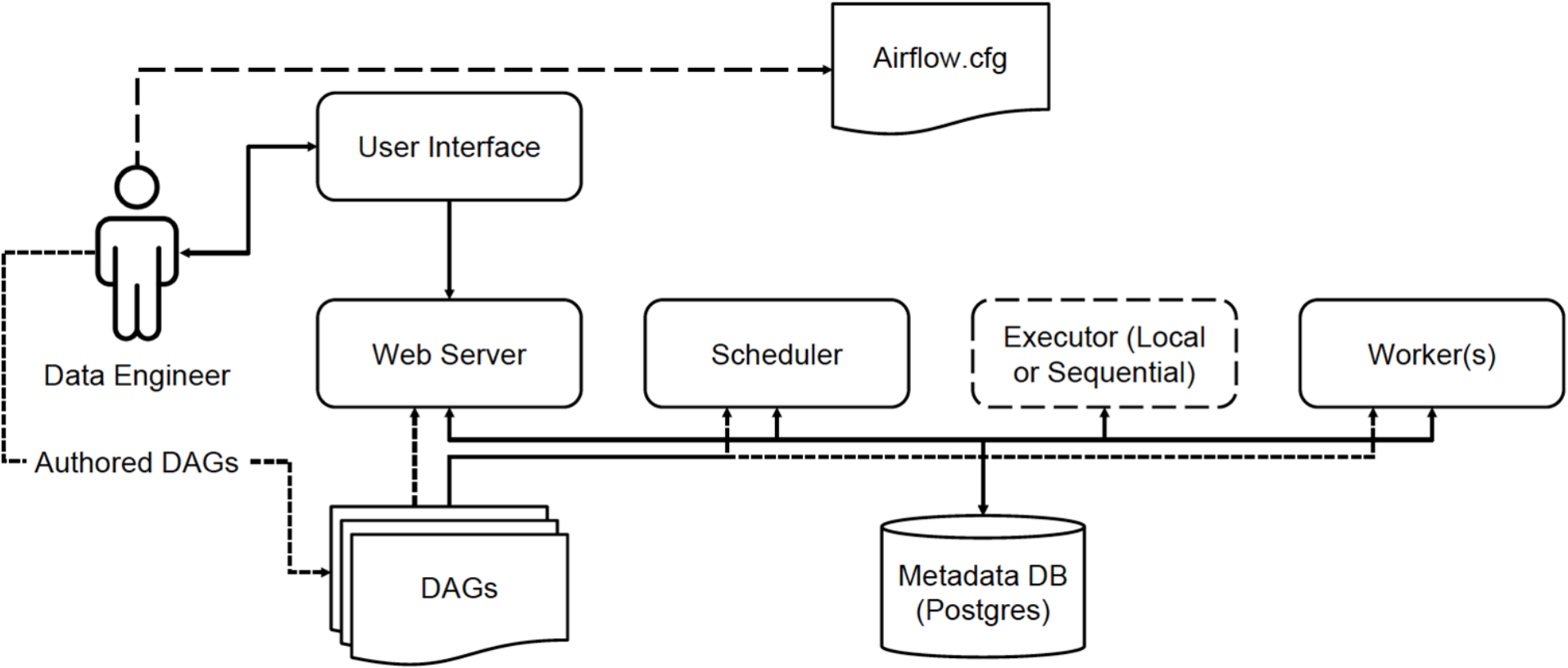
Архитектура



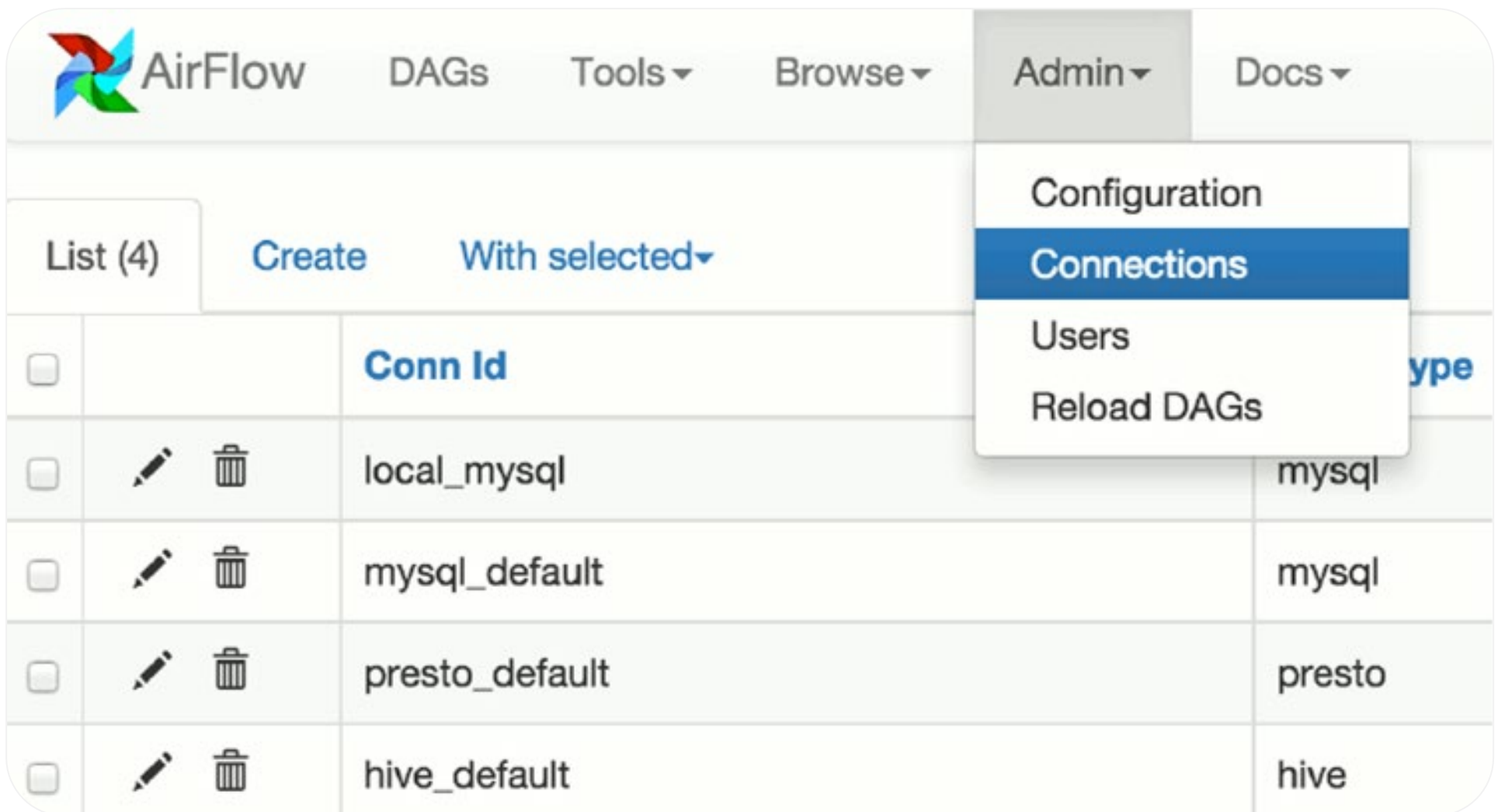
Архитектура











Архитектура



Больше возможностей



The screenshot shows the Apache AirFlow Admin interface. At the top, there is a navigation bar with the AirFlow logo and several tabs: DAGs, Tools, Browse, Admin, and Docs. The 'Admin' tab is currently selected, and its dropdown menu is open, showing options: Configuration, Connections (highlighted in blue), Users, and Reload DAGs. Below the navigation bar, there is a section for 'List (4)' with buttons for 'Create' and 'With selected'. Below this is a table with columns for 'Conn Id' and 'type'. The table contains four rows of connections: local_mysql, mysql_default, presto_default, and hive_default.

		Conn Id	type
<input type="checkbox"/>	 	local_mysql	mysql
<input type="checkbox"/>	 	mysql_default	mysql
<input type="checkbox"/>	 	presto_default	presto
<input type="checkbox"/>	 	hive_default	hive

- Переменные (Variables)
- Пулы (Pools)
- Хуки (Hooks)
- Подключения (Connections)
- Плагины (Plugins)
- ...

Выводы

- 1 Разобрали цели и задачи Airflow, причины его использования
- 2 Изучили основные понятия Airflow — DAG, scheduler, operator, task и др.
- 3 Рассмотрели архитектуру: веб-сервер, планировщик и т. д.