Оглавление

Содержание

Предисловие

Нам обоим посчастливилось работать инженерами по обработке данных в интересные и трудные времена.

Хорошо это или плохо, но многие компании и организации осознают, что данные играют ключевую роль в управлении их операциями и их улучшении. Последние разработки в области машинного обучения и искусственного интеллекта открыли множество новых возможностей для извлечения выгоды.

Однако внедрение процессов, ориентированных на данные, часто бывает затруднительным, поскольку обычно требуется координировать работу в различных гетерогенных системах и связывать все воедино аккуратно и своевременно для последующего анализа или развертывания продукта.

В 2014 году инженеры Airbnb осознали проблемы управления сложными рабочими процессами данных внутри компании. Чтобы разрешить их, они приступили к разработке Airflow: решения с открытым исходным кодом, позволяющего писать и планировать рабочие процессы, а также отслеживать их выполнение с помощью встроенного веб-интерфейса.

Успех этого проекта быстро привел к тому, что его приняли в рамках Apache Software Foundation, сначала в качестве проекта инкубатора в 2016 году, а затем в качестве проекта верхнего уровня в 2019 году. В результате многие крупные компании теперь используют Airflow для управления многочисленными критически важными процессами обработки данных.

Работая консультантами в GoDataDriven, мы помогли клиентам внедрить Airflow в качестве ключевого компонента в проектах, связанных с созданием озер и платформ данных, моделей машинного обучения и т. д. При этом мы поняли, что передача этих решений может быть непростой задачей, поскольку такие сложные инструменты, как Airflow, трудно освоить в одночасье. По этой причине мы также разработали программу по обучению Airflow в GoData-Driven, часто

организовывали семинары и участвовали в них, чтобы поделиться своими знаниями, мнениями и даже пакетами с открытым исходным кодом. В совокупности эти усилия помогли нам изучить тонкости работы с Airflow, которые не всегда было легко понять, используя доступную нам документацию.

В этой книге мы хотим предоставить исчерпывающее введение в Airflow, которое охватывает все, от построения простых рабочих процессов до разработки собственных компонентов и проектирования / управления развертываниями Airflow. Мы намерены дополнить блоги и другую онлайн-документацию, объединив несколько тем в одном месте в кратком и понятном формате. Тем самым мы надеемся придать вам стимул в работе с Airflow, опираясь на опыт, который мы накопили, и преодолевая трудности, с которыми мы столкнулись за последние годы.

Благодарности

Выход этой книги был бы невозможен без поддержки множества замечательных людей. Коллеги из GoDataDriven и наши друзья поддерживали нас и предоставили ценные предложения и важные наблюдения. Кроме того, читатели, члены программы Manning Early Access Program (MEAP), оставили полезные комментарии на онлайн-форуме.

Рецензенты, участвовавшие в процессе подготовки издания, также предоставили полезные отзывы: Эл Кринкер, Клиффорд Тербер, Дэниел Ламблин, Дэвид Криф, Эрик Платон, Фелипе Ортега, Джейсон Рендель, Джереми Чен, Джири Пик, Джонатан Вуд, Картик Сирасанагандла, Кент Р. Спиллнер, Лин Чен, Филип Бест, Филип Паттерсон, Рамбабу Поза, Ричард Мейнсен, Роберт Г. Гимбель, Роман Павлов, Сальваторе Кампанья, Себастьян Пальма Мардонес, Торстен Вебер, Урсин Стаусс и Влад Навицкий.

Мы выражаем особую благодарность Брайану Сойеру из издательства Manning, редактору отдела закупок, который помог нам сформировать первоначальную заявку на издание этой книги и поверил, что мы сможем довести дело до конца; Трише Лувар, нашему редактору-консультанту по аудитории, которая очень терпеливо отвечала на все наши вопросы, давала важные отзывы по каждой из наших черновиков глав и была для нас важным проводником на протяжении всего пути; а также остальным сотрудникам: редактору проекта Дейрдре Хиам; редактору Мишель Митчелл; корректору Кери Хейлз и техническому корректору Элу Кринкеру.

Бас Харенслак

Я хотел бы поблагодарить своих друзей и семью за их терпение и поддержку в течение этого приключения, длившегося полтора года, которое превратилось из второстепенного проекта в бесчисленное количество дней, ночей и выходных. Стефани, спасибо за то, что все время терпела меня, пока я работал за компьютером. Мириам, Герд и Лотте,

спасибо за то, что терпели меня и верили в меня, пока я писал эту книгу. Я также хотел бы поблагодарить команду GoDataDriven за их поддержку и стремление всегда учиться и совершенствоваться. Пять лет назад я и представить себе не мог, когда начал работать, что стану автором книги.

Джулиан де Руйтер

Прежде всего я хотел бы поблагодарить свою жену Анн Полин и сына Декстера за их бесконечное терпение в течение многих часов, которые я потратил на то, чтобы «еще немного поработать» над книгой. Эта книга была бы невозможна без их непоколебимой поддержки. Также хотел бы поблагодарить нашу семью и друзей за их поддержку и доверие. Наконец, хочу сказать спасибо нашим коллегам из GoDataDriven за их советы и поддержку, от которых я также многому научился за последние годы.

О книге

Эта книга была написана, чтобы помочь вам реализовать рабочие процессы (или конвейеры), ориентированные на обработку данных с помощью Airflow. Она начинается с концепций и механики, участвующих в программном построении рабочих процессов для Apache Airflow с использованием языка программирования Python. Затем мы переходим к более подробным темам, таким как расширение Airflow путем создания собственных компонентов и всестороннего тестирования рабочих процессов. Заключительная часть книги посвящена проектированию и управлению развертывания Airflow, затрагивая такие темы, как безопасность и проектирование архитектур для облачных платформ.

Кому адресована эта книга

Эта книга написана для специалистов и инженеров по обработке данных, которые хотят разрабатывать базовые рабочие процессы в Airflow, а также для инженеров, интересующихся более сложными темами, такими как создание собственных компонентов для Airflow или управление развертываниями Airflow. Поскольку рабочие процессы и компоненты Airflow построены на языке Python, мы ожидаем, что у читателей имеется промежуточный опыт программирования на Python (т. е. они хорошо умеют создавать функции и классы Python, имеют представление о таких понятиях, как * args и ** kwargs и т. д.). Также будет полезен опыт работы с Docker, поскольку большинство примеров кода запускаются с использованием Docker (хотя при желании их можно запускать локально).

Структура книги

Книга состоит из четырех разделов, которые охватывают 18 глав.

Часть I посвящена основам Airflow. В ней объясняется, что такое Airflow, и изложены его основные концепции:

- в главе 1 обсуждается концепция рабочих процессов / конвейеров обработки данных и их создание с помощью Apache Airflow. В ней также рассказывается о преимуществах и недостатках Airflow по сравнению с другими решениями, в том числе приводятся ситуации, при которых вы, возможно, не захотите использовать Apache Airflow;
- глава 2 посвящена базовой структуре конвейеров в Apache Airflow (также известных как ОАГ), с объяснением различных задействованных компонентов и их совместимости;
- в главе 3 показано, как использовать Airflow для планирования работы конвейеров через повторяющиеся промежутки времени, чтобы можно было (например) создавать конвейеры, которые постепенно загружают новые данные с течением времени. В этой главе также рассматриваются тонкости механизма планирования Airflow, который часто является источником путаницы;
- в главе 4 показано, как использовать механизмы создания шаблонов в Airflow для динамического включения переменных в определения конвейера. Это позволяет ссылаться на такие вещи, как даты выполнения расписания в конвейерах;
- в главе 5 демонстрируются различные подходы к определению отношений между задачами в конвейерах, что позволяет создавать более сложные структуры конвейеров с ветками, условными задачами и общими переменными.

В части II подробно рассматривается использование более сложных тем, включая взаимодействие с внешними системами, создание собственных компонентов и разработку тестов для ваших конвейеров:

- в главе 6 показано, как запускать рабочие процессы другими способами, не связанными с фиксированным расписанием, такими как загрузка файлов или вызов по протоколу HTTP;
- в главе 7 демонстрируются рабочие процессы с использованием операторов, которые управляют различными задачами вне Airflow, что позволяет создавать поток событий через системы, которые не связаны между собой;
- в главе 8 объясняется, как создавать собственные компоненты для Airflow, которые позволяют повторно использовать функциональные возможности в разных конвейерах или интегрироваться с системами, не поддерживающими встроенные функции Airflow;
- в главе 9 обсуждаются различные варианты тестирования рабочих процессов Airflow, затрагиваются свойства операторов и способы их применения во время тестирования;

- в главе 10 показано, как использовать рабочие процессы на базе контейнеров для выполнения задач конвейера в Docker или Kubernetes, а также обсуждаются преимущества и недостатки этих подходов.

Часть III посвящена применению Airflow на практике и затрагивает такие темы, как передовые методы, запуск и обеспечение безопасности Airflow и последний демонстрационный пример:

- в главе 11 рассказывается о передовых методах построения конвейеров, которые помогут вам разрабатывать и внедрять эффективные и удобные в сопровождении решения;
- в главе 12 подробно описано несколько тем, которые необходимо учитывать при запуске Airflow в промышленном окружении, например архитектуры для горизонтального масштабирования, мониторинга, журналирования и оповещений;
- в главе 13 обсуждается, как обезопасить установку Airflow, чтобы избежать нежелательного доступа и минимизировать воздействие в случае взлома;
- в главе 14 показан пример проекта Airflow, в котором мы периодически обрабатываем поездки с использованием сервисов Yellow Cab и Citi Bikes по Нью-Йорку, чтобы определить самый быстрый способ передвижения между районами.

В части IV исследуется, как запускать Airflow в облачных платформах. Она включает такие темы, как проектирование развертываний Airflow для облачных платформ и использование встроенных операторов для взаимодействия с облачными сервисами:

- в главе 15 дается общее введение, в котором рассказывается, какие компоненты Airflow участвуют в (облачных) развертываниях, представлена идея, лежащая в основе облачных компонентов, встроенных в Airflow, и рассматриваются варианты самостоятельной реализации развертывания в облаке в сравнении с использованием управляемого решения;
- глава 16 посвящена облачной платформе Amazon AWS. Здесь рассказывается о решениях для развертывания Airflow на AWS и демонстрируется, как применять определенные компоненты для использования сервисов AWS;
- в главе 17 разрабатываются развертывания и демонстрируются облачные компоненты для платформы Microsoft Azure;
- в главе 18 рассматриваются развертывания и облачные компоненты для платформы Google GCP.

Тем, кто плохо знаком с Airflow, следует прочитать главы 1 и 2, чтобы получить внятное представление о том, что такое Airflow, и его возможностях. В главах 3–5 представлена важная информация об основных функциях Airflow. В остальной части книги обсуждаются такие темы, как создание собственных компонентов, тестирование, передовые практики и развертывание, и ее можно читать в произвольном порядке в зависимости от конкретных потребностей читателя.

О коде

Весь исходный код в листингах или тексте набран моноширинным шрифтом, чтобы отделить его от обычного текста. Иногда также используется **жирный шрифт**, чтобы выделить код, который изменился по сравнению с предыдущими шагами в главе, например когда к существующей строке кода добавляется новая функция.

Во многих случаях оригинальный исходный код был переформатирован; мы добавили разрывы строк и переработали отступы, чтобы уместить их по ширине книжных страниц. В редких случаях, когда этого оказалось недостаточно, в листинги были добавлены символы продолжения строки (➥). Кроме того, комментарии в исходном коде часто удаляются из листингов, когда описание кода приводится в тексте. Некоторые листинги сопровождают аннотации, выделяющие важные понятия.

Ссылки на элементы в коде, сценарии или определенные классы, переменные и значения Airflow часто выделяются *курсивом*, чтобы их было легче отличить от окружающего текста.

Исходный код всех примеров и инструкции по их запуску с помощью Docker и Docker Compose доступны в нашем репозитории GitHub (https://github.com/BasPH/data-pipelines-with-apache-airflow), а также на сайте издательства «ДМК Пресс» www.dmkpress.com на странице с описанием соответствующей книги.

> **ПРИМЕЧАНИЕ** В приложении A представлены более подробные инструкции по запуску примеров кода.

Все примеры кода были протестированы с использованием Airflow 2.0. Большинство примеров также следует запускать в более старых версиях Airflow (1.10) с небольшими изменениями. Там, где это возможно, мы включили встроенные указатели, как это сделать. Чтобы вам было легче учитывать различия в путях импорта между Airflow 2.0 и 1.10, в приложении B представлен обзор измененных путей импорта.

Отзывы и пожелания

Мы всегда рады отзывам наших читателей. Расскажите нам, что вы думаете об этой книге, – что понравилось или, может быть, не понравилось. Отзывы важны для нас, чтобы выпускать книги, которые будут для вас максимально полезны.

Вы можете написать отзыв на нашем сайте www.dmkpress.com, зайдя на страницу книги и оставив комментарий в разделе «Отзывы и рецензии». Также можно послать письмо главному редактору по адресу dmkpress@gmail.com; при этом укажите название книги в теме письма.

Если вы являетесь экспертом в какой-либо области и заинтересованы в написании новой книги, заполните форму на нашем сайте по адресу http://dmkpress.com/authors/publish_book/ или напишите в издательство по адресу dmkpress@gmail.com.

Список опечаток

Хотя мы приняли все возможные меры для того, чтобы обеспечить высокое качество наших текстов, ошибки все равно случаются. Если вы найдете ошибку в одной из наших книг, мы будем очень благодарны, если вы сообщите о ней главному редактору по адресу dmkpress@gmail.com. Сделав это, вы избавите других читателей от недопонимания и поможете нам улучшить последующие издания этой книги.

Нарушение авторских прав

Пиратство в интернете по-прежнему остается насущной проблемой. Издательства «ДМК Пресс» и Manning Publications очень серьезно относятся к вопросам защиты авторских прав и лицензирования. Если вы столкнетесь в интернете с незаконной публикацией какой-либо из наших книг, пожалуйста, пришлите нам ссылку на интернет-ресурс, чтобы мы могли применить санкции.

Ссылку на подозрительные материалы можно прислать по адресу электронной почты dmkpress@gmail.com.

Мы высоко ценим любую помощь по защите наших авторов, благодаря которой мы можем предоставлять вам качественные материалы.

Об авторах

БАС ХАРЕНСЛАК – инженер по обработке данных в GoDataDriven, компании, занимающейся разработкой решений на основе данных, расположенной в Амстердаме, Нидерланды. Имея опыт разработки программного обеспечения и информатики, он любит работать над программным обеспечением и данными, как если бы они были сложными головоломками. Он предпочитает работать над ПО с открытым исходным кодом, участвует в проекте Apache Airflow и является одним из организаторов семинара Amsterdam Airflow.

ДЖУЛИАН ДЕ РУИТЕР – инженер машинного обучения, имеет опыт работы в области компьютерных наук и наук о жизни, а также имеет докторскую степень в области вычислительной биологии рака. Будучи опытным разработчиком программного обеспечения, он любит соединять миры науки о данных и инженерии, используя облачное программное обеспечение и программное обеспечение с открытым исходным кодом для разработки решений машинного обучения, готовых к промышленной эксплуатации. В свободное время он с удовольствием разрабатывает собственные пакеты Python, участвует в проектах с открытым исходным кодом и возится с электроникой.

Об иллюстрации на обложке

Рисунок на обложке называется «Femme de l'Isle de Siphanto», или *Женщина с острова Сифанто*. Иллюстрация взята из коллекции костюмов разных стран Жака Грассе де Сен-Совера (1757–1810), «Costumes de Différents Pays», изданной во Франции в 1797 году.

Каждая иллюстрация нарисована и раскрашена вручную. Богатое разнообразие коллекции Грассе де Сен-Совера наглядно напоминает нам о том, насколько обособленными в культурном отношении были города и регионы мира всего 200 лет назад. Изолированные друг от друга люди говорили на разных диалектах и языках. На улице или в деревне было легко определить, где они живут и чем занимаются или каково их положение в обществе, по их одежде.

С тех пор наша манера одеваться изменилась, а разнообразие регионов, столь богатых в то время, исчезло. Сейчас трудно отличить жителей разных континентов, не говоря уже о разных городах, регионах или странах. Возможно, если смотреть на это оптимистично, мы обменяли культурное и визуальное разнообразие на более разнообразную частную жизнь или более разнообразную интеллектуальную и техническую жизнь.

Мы в Manning высоко ценим изобретательность, инициативу и, конечно, радость от компьютерного бизнеса с книжными обложками, основанными на разнообразии жизни в разных регионах два века назад, которое оживает благодаря рисункам Грассе де Сен-Совера.