

АВТОМАТИЗАЦИЯ СБОРА И ОБРАБОТКИ ДАННЫХ УЗЛОВ И ДЕТАЛЕЙ ЖЕЛЕЗНОДОРОЖНОГО ПОДВИЖНОГО СОСТАВА С ИСПОЛЬЗОВАНИЕМ AIRFLOW ДЛЯ МАШИННОГО ОБУЧЕНИЯ



Н.О. Мироненко



О.И. Мироненко

В статье рассматривается применение Airflow для автоматизации управления данными железнодорожного транспорта с целью повышения точности и оперативности анализа состояния подвижного состава. Исследование направлено на демонстрацию преимуществ внедрения Airflow в процесс обработки данных и его влияние на качество прогнозирования состояния узлов и деталей вагонов.

Ключевые слова: разработка технологических процессов, ремонтное производство, оптимизация транспортных операций

EDN: ORJCMB

Основное внимание цифровизации на транспорте в рамках Индустрии 4.0 и Транспортной стратегии Российской Федерации до 2030 года направлено на использование цифровых технологий автоматизации рабочих процессов и повышение точности анализа состояния подвижного состава с использованием искусственного интеллекта. В настоящее время использование цифровых технологий еще недостаточно широко используется на предприятиях. Вместо этого используются бумажные носители, файлы программы Microsoft Excel и т.д., затрудняя автоматизацию технологических процессов, что приводит к увеличению времени выполнения задач и к переносу их срока выпол-

нения. При этом сотрудникам необходимо выполнять однотипные задачи, не требующие интеллектуальной деятельности. Это в разы увеличивает ошибки из-за человеческого фактора. К самым частым ошибкам можно отнести: пропуски, дублирование, ошибки в составлении однотипных отчетов и т.д. Для решения этих проблем в статье рассматривается система Apache Airflow (Airflow), которая позволяет автоматизировать однотипные задачи, что необходимо для получения корректных данных для машинного обучения [1–3].

Airflow представляет собой платформу, которая обеспечивает:

- сбор данных;

Мироненко Никита Олегович, студент Московского государственного технического университета имени Н.Э. Баумана (национальный исследовательский университет). Область научных интересов: машинное обучение, нейронные сети, анализ данных в задачах железнодорожного транспорта, системное администрирование, разработка СДО. Автор 46 научных работ. Имеет восемь патентов на изобретения.

Мироненко Олег Игоревич, кандидат технических наук, доцент кафедры «Нетяговый подвижной состав» Российского университета транспорта (РУТ (МИИТ)). Область научных интересов: динамика систем, организация вагонной отрасли, машинное обучение, нейронные сети, разработка СДО, системное администрирование. Автор 84 научных работ. Имеет семь патентов на изобретения.

Сергеев Константин Александрович, доктор технических наук, доцент, доцент кафедры «Нетяговый подвижной состав» Российского университета транспорта (РУТ (МИИТ)). Область научных интересов: технологическая подготовка производства вагоноремонтных предприятий. Автор 142 научных работ, в том числе двух монографий.

Бомбардиров Андрей Петрович, кандидат технических наук, доцент кафедры «Нетяговый подвижной состав» Российского университета транспорта (РУТ (МИИТ)). Область научных интересов: динамика и прочность нетягового подвижного состава (вагонов), технологическая подготовка производства вагоноремонтных предприятий. Автор 136 научных работ. Имеет три патента на изобретения.

- планирование задач;
- организацию рабочих процессов;
- обработку ошибок;
- мониторинг выполнения;
- интеграцию с другими системами;
- логирование;
- работу с базой данных;
- параллельное выполнение задач;
- уведомление об успешном или неудачном завершении события;
- визуализацию процессов выполнения.

Следует отметить, что Airflow предоставляет возможность гибкого и быстрого интегрирования в производственные процессы, позволяя автоматизировать сбор данных из различных источников вагонных предприятий ОАО «РЖД», сохраняя их в единой информационной системе, обеспечивая оперативность анализа и последующего прогнозирования технического состояния узлов и деталей вагонов для своевременного выявления потенциальных неисправностей и для принятия соответствующих мер для их устранения [1–5].

Рассмотрим элементы для функционирования Airflow в ОАО «РЖД»:

1. Ациклический граф.
2. Airflow в контейнерах.
3. Интерфейс Airflow.
4. Планирование задач.
5. Обработка пользовательских событий.
6. Запуск через равные промежутки времени.
7. Интервалы на основе Сноп.
8. Инкрементная обработка.

9. Атомарность.
10. Идемпотентность.

Ациклический граф

Концепция Airflow состоит из работы связанных задач, которые могут выполняться как последовательно, так и параллельно. К задачам в рамках ОАО «РЖД» можно отнести: получение данных из различных депо, обработку полученных данных, приведение данных в нужный формат и т.д. Для описания этих задач применяется ориентированный ациклический граф (DAG). На рис. 1 показаны задачи, которые выполняются в определенной последовательности:

1. Получение данных.
2. Уведомление о получении данных.
3. Валидация данных.
4. Уведомление о валидации данных.
5. Приведение в нужный формат.
6. Уведомление о приведении в нужный формат.
7. Обработка полученных данных.
8. Уведомление об обработке полученных данных.
9. Подключение к базе данных.
10. Перестройка индексов базы данных.
11. Сохранение в базе данных.
12. Уведомление о сохранении в базе данных.

Запуск Airflow в контейнерах Docker

Одним из преимуществ Airflow является возможность его запуска внутри контейнера Docker. Контейнеризация Docker позволяет: запускать приложение в изолированной среде операционной

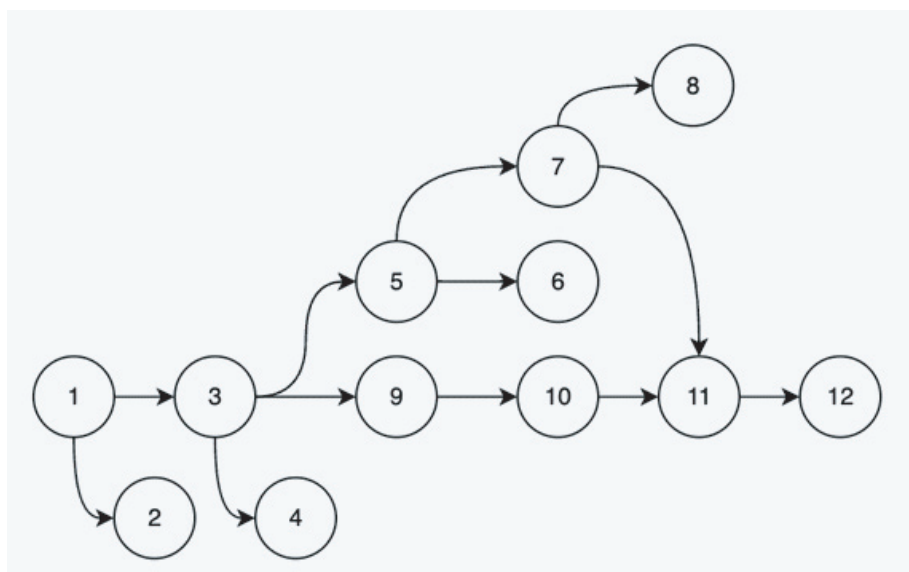


Рис. 1. Ациклический граф

системы, что увеличивает гибкость и надежность работы системы Airflow, производить быстрое масштабирование, одновременный запуск нескольких версий приложения на одном сервере, развертывание новых версий приложений, гибкое ограничение ресурсов, легкое интегрирование его с другими сервисами, управление различными версиями зависимостей библиотек и так далее (рис. 2).

Рабочее место оператора вагонного депо. Интерфейс Airflow

Взаимодействие с Airflow возможно при помощи нескольких интерфейсов, одним из которых является веб-интерфейс. Он упрощает управление и мониторинг задач, позволяя оператору отслеживать состояние и их выполнение, управлять параметрами запуска, анализировать логи и т.д. Это позволяет в случае непредвиденных ситуаций принять оператив-

ные меры для их решения, что важно на железнодорожных предприятиях.

На рис. 3 показан пример визуализации DAG в виде графа, который иллюстрирует процесс выполнения задач, где:

- темно-зеленым цветом показаны успешно выполненные задачи;
- ярко-зеленым цветом показано выполнение текущей задачи;
- красным цветом показана задача, выполненная с ошибкой.

Основное рабочее место оператора Airflow представлено на рис. 4. Рабочее место состоит из трех областей: в верхней области показывается статус текущего выполнения задач, левая область показывает исторические и текущий статусы выполнения, правая область отвечает за визуализацию DAG в виде графа или диаграммы Ганта.

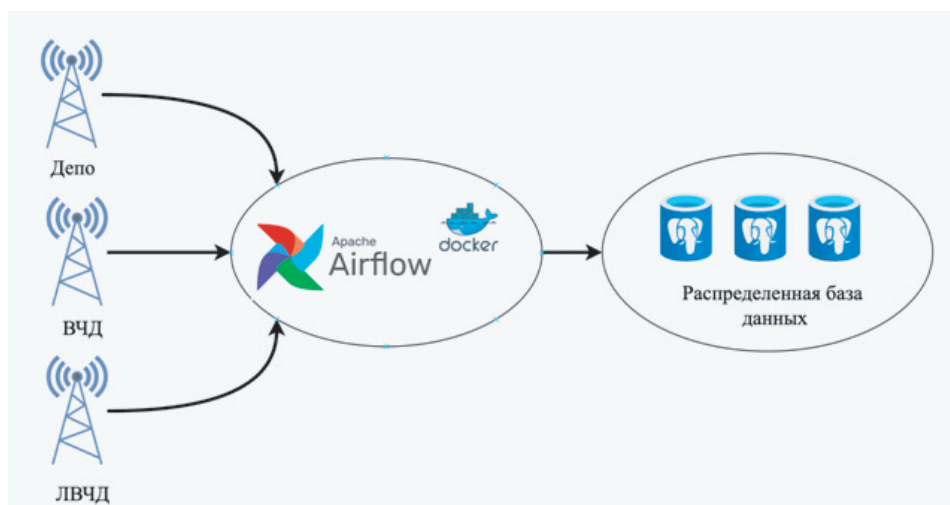


Рис. 2. Интеграция Airflow с другими сервисами

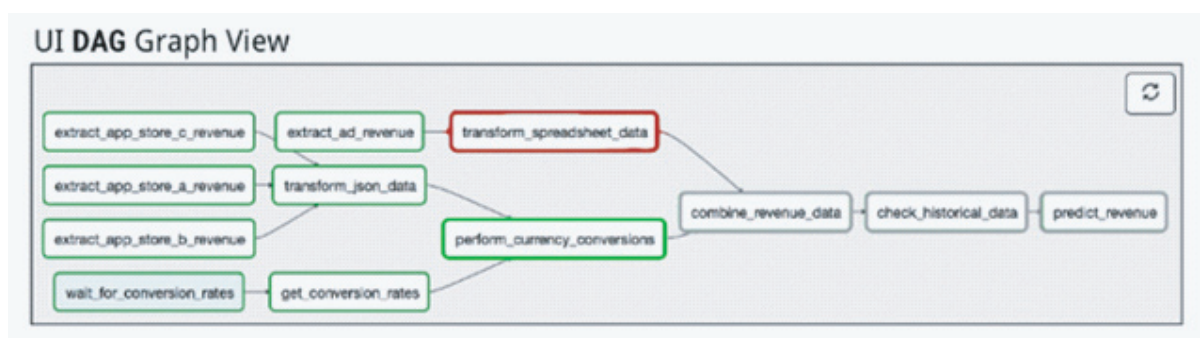


Рис. 3. Пользовательский интерфейс Airflow (Граф)

В левом окне показано историческое выполнение задач DAG, где отмечено следующее:

- Failed — ошибка выполнения;
- Queued — поставлена в очередь на выполнение;
- Removed — задача удалена;
- Restarting — перезапуск задачи;
- Running — задача выполняется;
- Skipped — пропущена;
- Success — успешное выполнение;
- Up_for_retry — готова к перезапуску.

В правой стороне показано окно выполнения задач графа на текущий момент. Отметим, что Airflow позволяет гибко настраивать пользовательский интерфейс выполняемых и выполненных задач с предоставлением информации в виде:

- детального представления;
- графа;
- диаграммы Ганта;
- исходного кода;
- журнала аудита;
- логов.

Планирование задач в Airflow

Одной из важных особенностей Airflow является гибкая возможность планирования и выполнения задач. Как пример, задачи по сбору данных настраиваются на выполнение каждые 5 минут, а задачи по анализу и прогнозированию — раз в день. Такая гибкость позволяет равномерно распределить нагрузку на систему и обеспечить своевременное получение аналитической информации, необходимой для принятия решений.

Данный подход включает в себя возможность запуска как задачи, так и совокупность задач внутри DAG по расписанию или вручную. Использование мануального режима позволяет запускать задачи, не дожидаясь выполнения других задач DAG, что может быть полезно при исследовании новых функциональных возможностей системы или при внештатных ситуациях, массовой рассылке сообщений сотрудникам предприятия и т.д.

Инкрементная обработка данных

В Airflow имеется функция инкрементной обработки данных (рис. 5), которая дает возможность

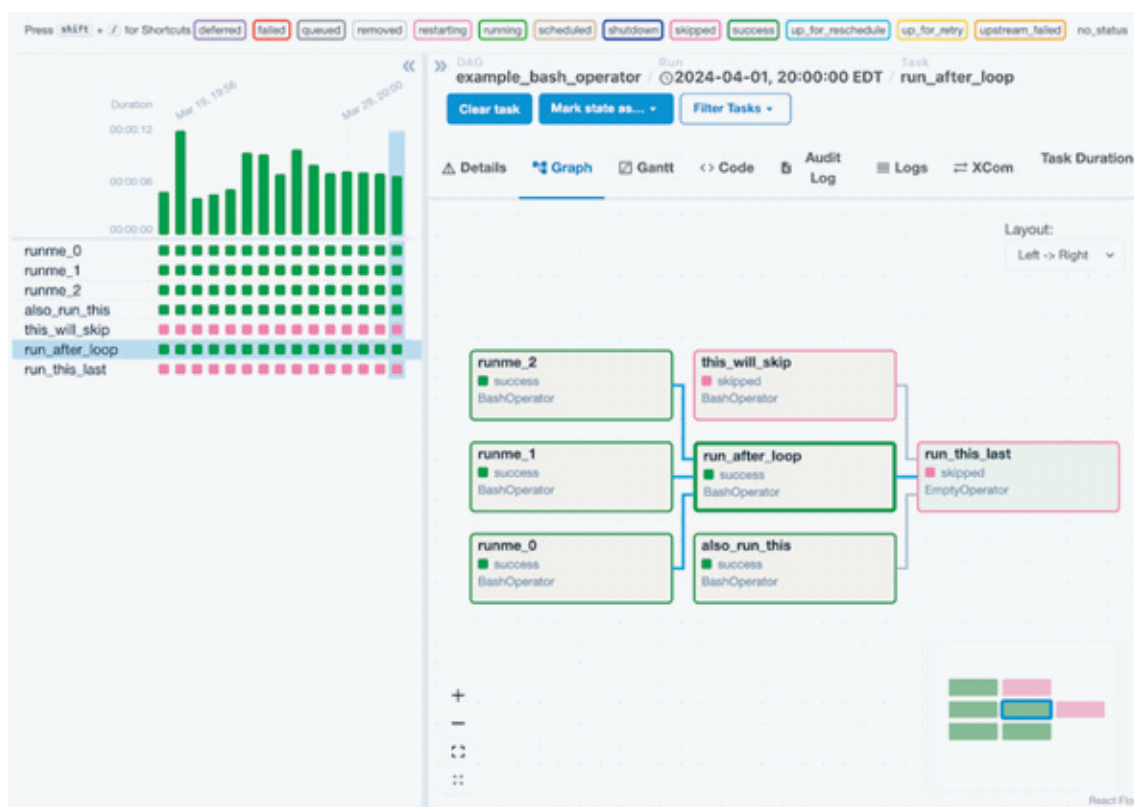


Рис. 4. Рабочее место оператора Airflow

обрабатывать только те данные, которые были изменены или добавлены с момента последней обработки, что позволяет значительно снизить нагрузку и ускорить процесс выполнения задач.

Атомарность

Атомарность задач обеспечивает целостность и завершенность выполнения операций в рамках системы. В Airflow это означает, что каждая задача должна быть выполнена до конца или не быть выполненной вообще, без частичного завершения. Это важно для обеспечения точности и надежности обработки данных, особенно в распределенных системах, где сбои недопустимы. Атомарность связана с транзакционностью (рис. 6), характерной для баз данных. В транзакционных системах операция должна

быть выполнена полностью или откатана, что предотвращает частичное выполнение и потенциальные несоответствия данных. В Airflow атомарность задач достигается с помощью различных механизмов управления состоянием задач и механизмов отката в случае ошибок.

Идемпотентность

Идемпотентность задач (рис. 7) в Airflow означает, что повторное выполнение задачи даст тот же результат, что и первое. Это свойство корректно для задач, связанных с обработкой данных, где возможно повторное выполнение задач из-за сбоев или других причин.

Задачи по сбору данных могут быть настроены таким образом, чтобы их повторное выполнение

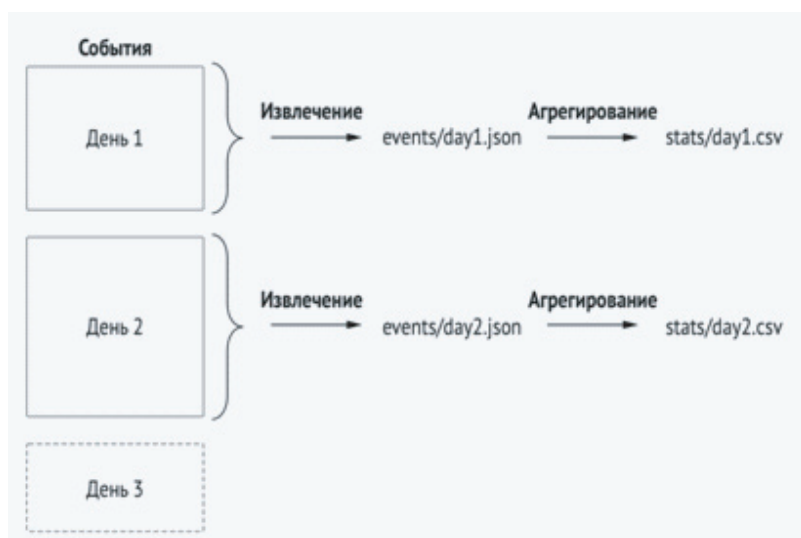


Рис. 5. Инкрементная обработка данных

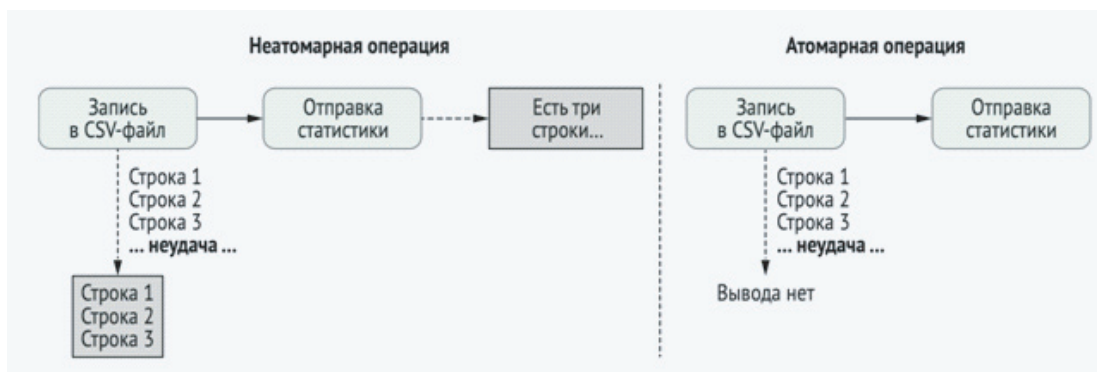


Рис. 6. Сравнение атомарной и неатомарной операции

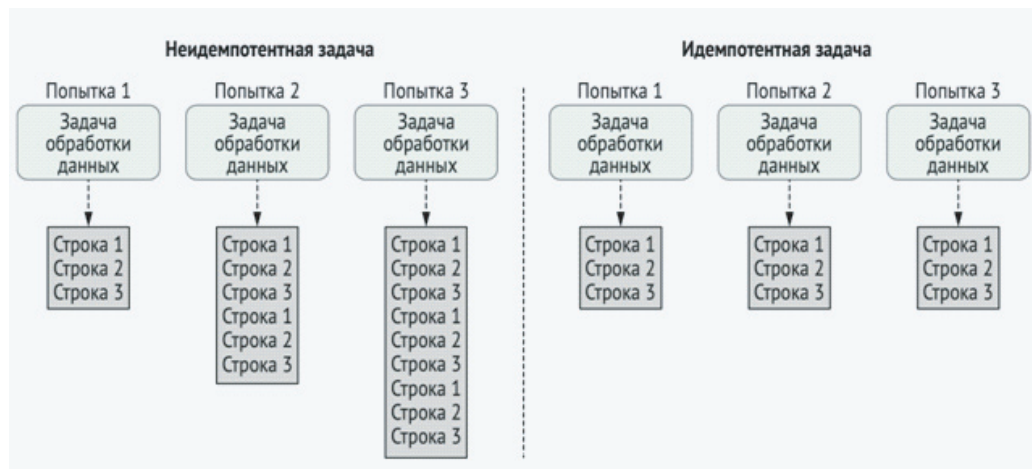


Рис. 7. Сравнение идемпотентной и неидемпотентной задачи

не приводило к дублированию данных, что обеспечивает корректность и целостность информации. Airflow предоставляет механизмы для реализации идемпотентных задач, что значительно упрощает управление процессами обработки данных.

Выводы

Применение Airflow в управлении данными железнодорожного транспорта позволит автоматизировать и повысить точность анализа. Использование DAG и механизмов обеспечения атомарности и идемпотентности задач улучшит качество обработки данных. Эти аспекты повысят надежность и эффективность прогнозирования технического состояния узлов и деталей подвижного состава железнодорожного транспорта. 

Литература

1. Плас, Джейк Вандер. Python для сложных задач. Наука о данных и машинное обучение. Руководство / Плас Джейк Вандер. - Москва : Питер, 2018. - 759 с. - Текст : непосредственный.
2. Taweh Beysolow II. Applied Reinforcement Learning with Python: With OpenAI Gym, Tensorflow, and Keras. Apress, 2019. - ISBN 1484251261. - Текст : непосредственный.
3. Колесникова, Г. И. Искусственный интеллект: проблемы и перспективы / Г. И. Колесникова. - Текст : электронный // Видеонаука: сетевой журнал. - 2018. - № 2 (10). - URL: <https://videonauka.ru/stati/44-novye-tehnologii/190-iskusstvennyj-intellekt-problemy-i-perspektivy> (дата обращения: 20.05.2025).
4. Солнцева, О. Г. Аспекты применения технологий искусственного интеллекта / О. Г. Солнцева. - Текст : электронный // E-Management. - 2018. - № 1. - URL: <https://cyberleninka.ru/article/n/aspekty-primeneniya-tehnologii-iskusstvennogo-intellekta> (дата обращения: 20.05.2025).
5. Шарден, Б. Крупномасштабное машинное обучение вместе с Python: учебное пособие / Б. Шарден, Л. Массарон, А. Боскетти; перевод с английского А. В. Логунова. - Москва : ДМК Пресс, 2018. - 358 с. - ISBN 978-5-97060-506. - Текст : непосредственный.