РТУ МИРЭА, Институт Информационных Технологий, Кафедра Прикладной Математики

Обзор Knime Analytics Platform

Автор: Артём Александрович Московка, ИКБО-20-19

Москва, 2022

Подробнее о KNIME

- Knime Analytics Platform – open-source фреймворк для анализа данных. Данный фреймворк позволяет реализовывать полный цикл анализа данных, включающий чтение данных из различных источников, преобразование и фильтрацию, собственно анализ, визуализацию и экспорт.
- Кому может быть интересна эта платформа:
  - Тем, кто хочет анализировать данные;
  - Тем, кто не владеет навыками программирования;
  - Тем, кто хочет покопаться в неплохой библиотеке реализованных алгоритмов и, возможно, узнать что-то новое.

Подробнее о KNIME

- Кому подходит: фрилансерам, малому и среднему бизнесу, ИП, специалистам, НКО, корпорациям.
- Развертывание: ПК, сервер предприятия, облако (SaaS).
- Работает на ОС: macOS, Windows, Linux.
- Стоимость: бесплатно, подписка для команд и организаций, особые условия для образования.
- Поддерживаемые языки: только английский.
- Свободное ПО.
- Имеется пробная версия и большая библиотека знаний.

Назначение системы KNIME Analytics Platform

Анализ данных (САД)

Интеллектуальный анализ данных (ИАД)

Machine Learning (ML)

Data Analysis (DA)

Big Data Analytics (BDA)

Data Mining (DM)

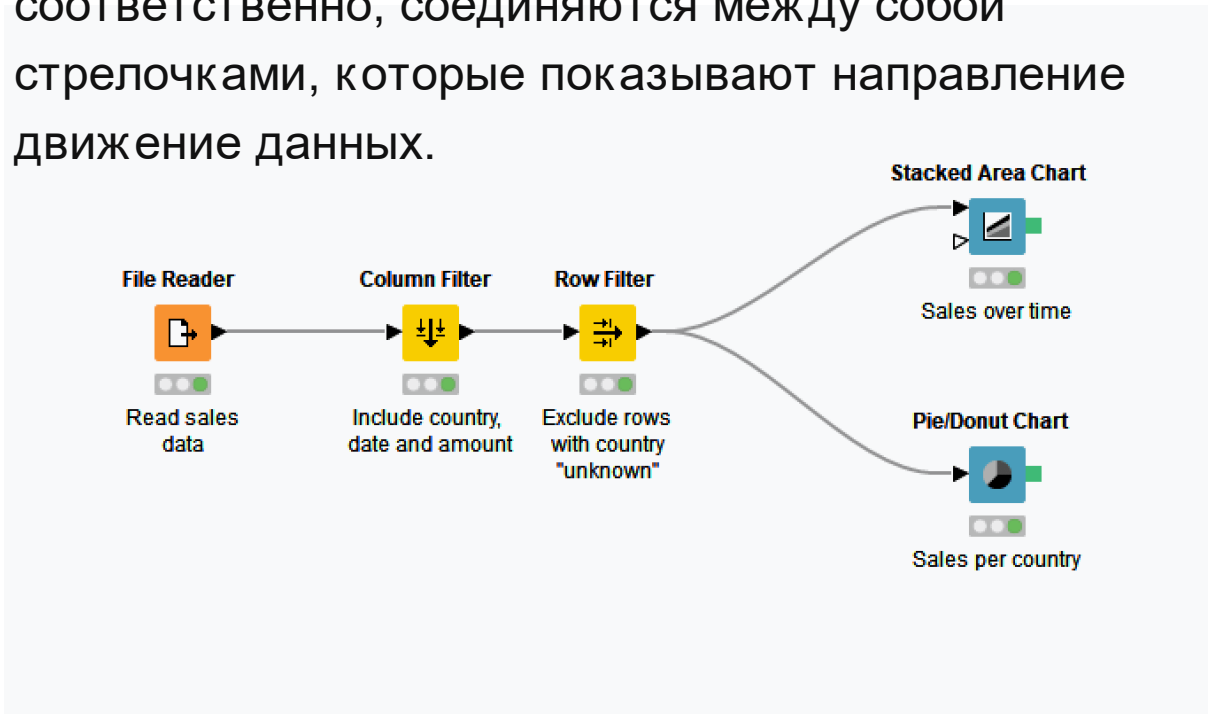
Predictive Analytics (PA)

Разработчик системы KNIME

- Страна разработки: Швейцария
- Головной офис: Цюрих, Швейцария
- Веб-сайт: https://www.knime.com/about

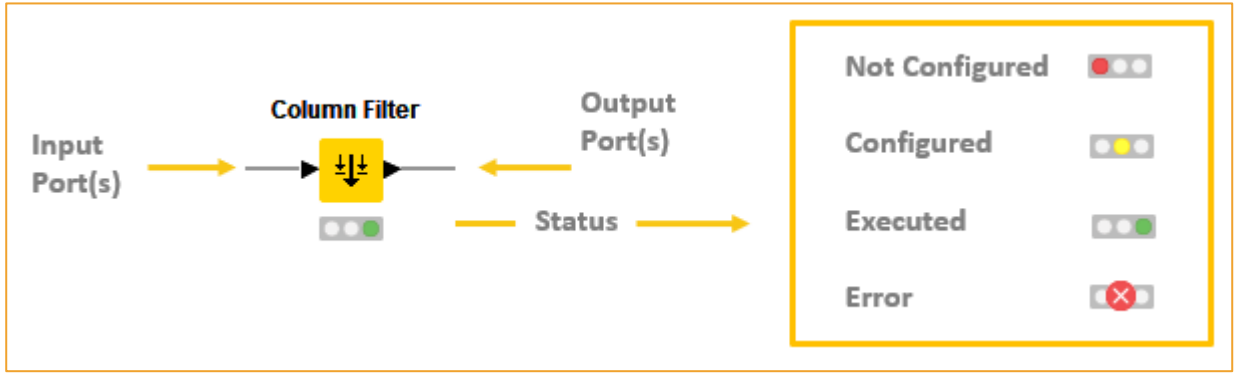
Workflows (рабочие потоки)

- В Knime процесс программирования логики осуществляется через создание рабочего потока. Workflow состоит из узлов, которые выполняют ту или иную функцию (например чтение данных из БД, трансформация, визуализация). Узлы, соответственно, соединяются между собой стрелочками, которые показывают направление движение данных.



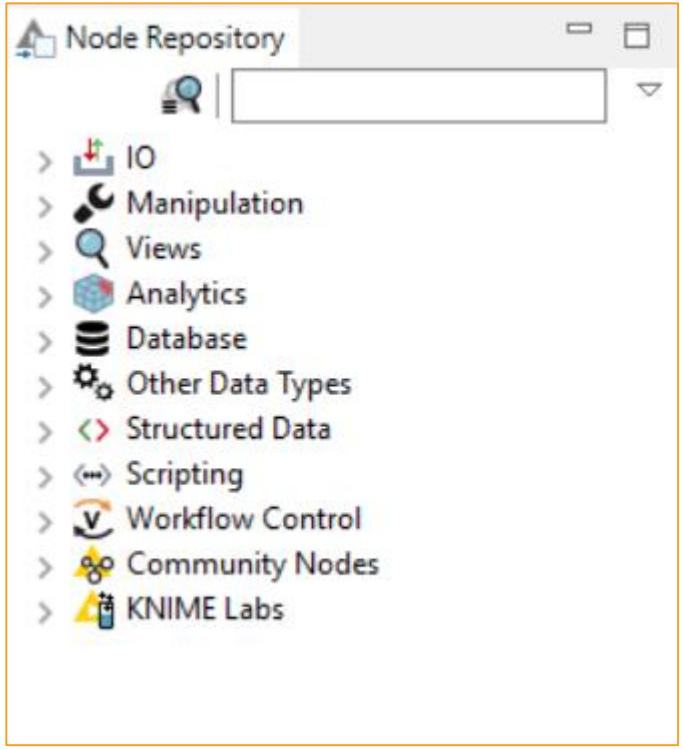
Workflows (рабочие потоки)

- После того, как workflow будет запущен на исполнение, в базовом сценарии узлы workflow начинают отрабатывать один за одним, начиная с самого первого. Если в ходе выполнения того или иного узла произошла ошибка, то исполнение всей ветки, следующей за ним, прекращается. Существует возможность перезапуска workflow не с первого, а с произвольного узла (debug mode).
- Светофор у каждого узла отражает его текущее состояние: красный – не настроен, желтый – готов к исполнению, зеленый – выполнен, крест – ошибка.



Nodes

- Workflow состоит из узлов (или «нод»). Практически у каждого узла есть конфигурационный диалог, в котором можно настраивать свойства.
- Все узлы разбиты на категории:

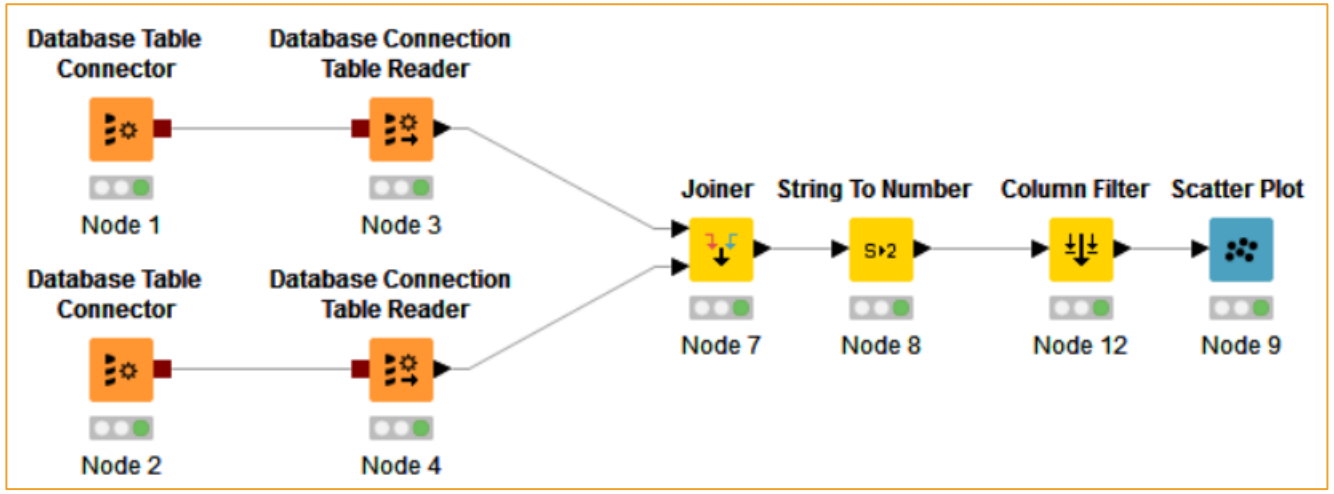


Nodes

- Поддерживается следующие типы узлов: IO — ввод/вывод данных (например чтение CSV), Manipulation – преобразование данных (включая фильтрацию строк, столбцов, сортировку), Views – визуализация данных (построение различных графиков включая Histogram, Pie Chart, Scatter Plot, etc.), Database – возможность подключения к базе данных, чтения/записи, Workflow Control – создание циклов, итерирование групп в ходе выполнения workflow и прочее.
- Из узлов реализующих анализ данных доступны различные статистические методы (включая линейную корреляцию, проверку гипотез), а также Data Mining методы (например нейронные сети, построение decision trees, cluster view).

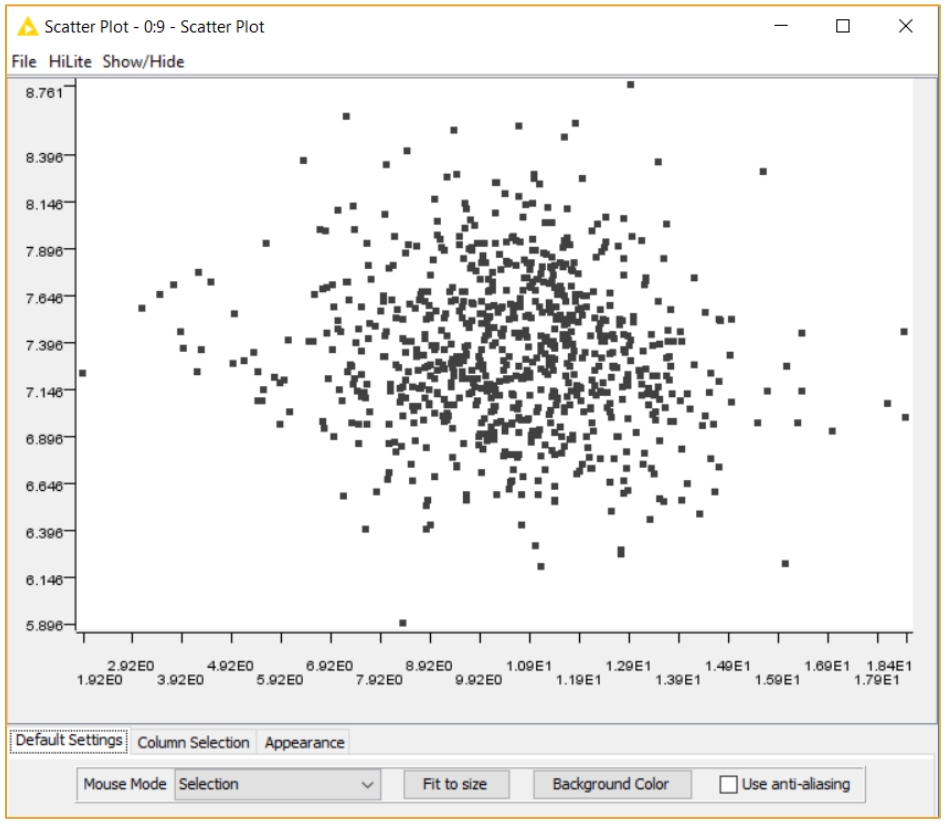
Пример workflow #1: построение простого Scatter Plot

- Рассмотрим пример простого workflow, который вытягивает данные, производит JOIN значений по некоему полю ID, фильтрацию и визуализацию результата на Scatter Plot.



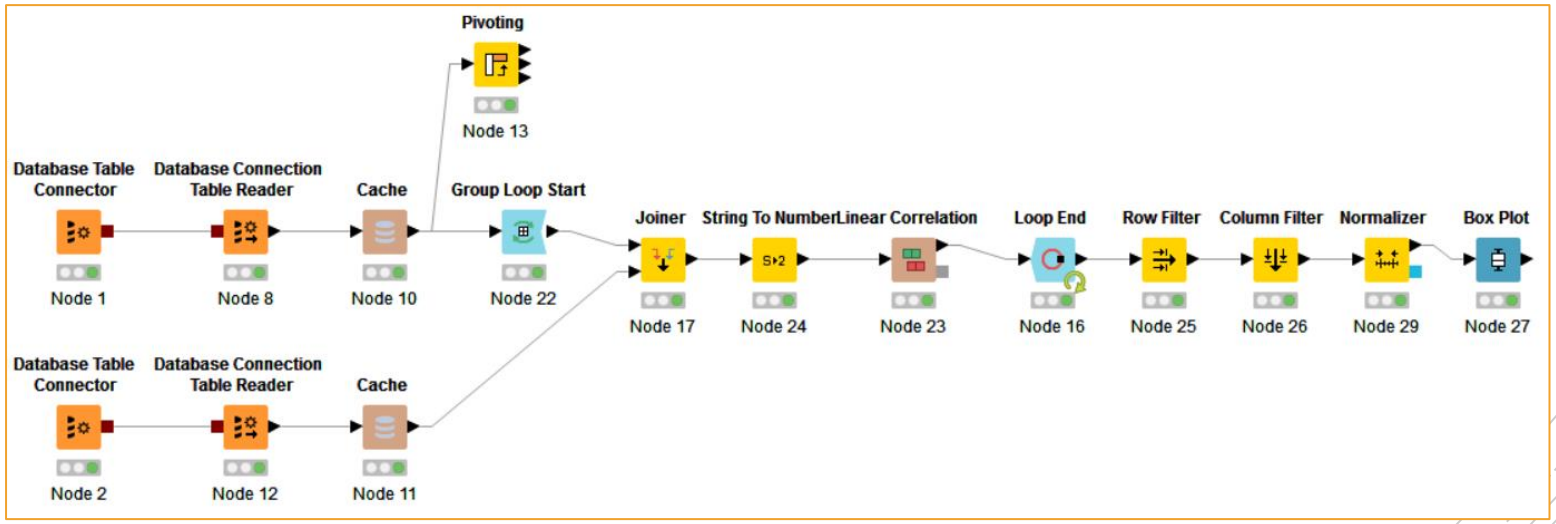
Пример workflow #1: построение простого Scatter Plot

- Построенный график открывается в новом окне.



Пример workflow #2: Correlation Analysis

- Рассмотрим еще один пример. Требуется сделать относительно большую выборку данных из БД, сгруппировать выборку по значениям некоего поля, и внутри каждой группы найти корреляцию значений из этой группы и целевого вектора.

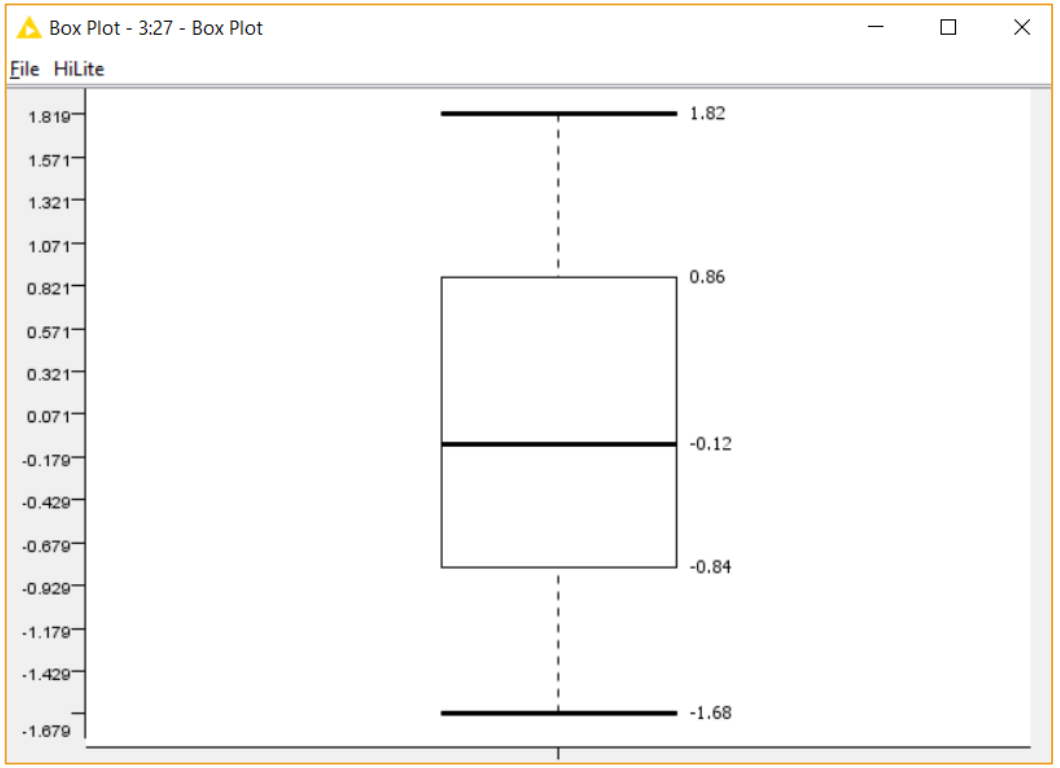


Пример workflow #2: Correlation Analysis

- В данном примере открывается два соединения к БД. Через одно соединение (Node 2) SQL-запросом вытягивается вектор из нескольких значений. Это будет целевой вектор, к которому будем искать корреляцию.
- Через другое соединение (также SQL-запросом) вытягивается относительно большая выборка данных. Далее данные попадают в Group Loop Start – Workflow Control оператор который делает GROUP BY, внутри этого loop-а к данным JOIN-ится целевой вектор, делается преобразование строковых значений в численные и считается линейная корреляция. Результаты вычислений аккумулируются в узле Loop End. На выходе из этого узла применяется фильтрация по строкам и столбцам, нормализация значений и построение Box Plot.

Пример workflow #2: Correlation Analysis

- После исполнения всего workflow и нажатия на View: Box Plot открывается окно с подсчитанными значениями для Box Plot.



Интересные возможности

- Исполнение workflow на сервере и предоставление доступа к результатам работы через REST API. Данная функциональность доступна при покупке KNIME-Server;
- Полный дистрибутив KNIME-а со всеми плагинами весит почти 2 гигабайта. В данный дистрибутив входит большое количество сторонних библиотек (например JFreeChart), которые становятся доступны в виде узлов;
- Реализована возможность сделать операцию Pivot прямо на базе данных или на данных, загруженных в локальный кеш;
- Доступна большая библиотека примеров;
- Работа с Hadoop и другими BigData источниками.

Выводы

- Данный фреймворк хорошо подойдет для людей, не сильно знакомых с программированием, с его помощью можно быстро создавать простые и средней сложности workflows и предоставлять к ним доступ через REST. Это может быть востребовано в каких-либо организациях.
- Data scientists, возможно, тоже найдут для себя много интересного и могут рассмотреть эту систему как дополнение к R или Python.
- Этот фреймворк хорош также для работы со студентами, поскольку наглядно видно все, что происходит с данными, по каким веткам они перемещаются и как преобразуются. Студенты могут изучать реализацию существующих узлов, дописывать свои компоненты (узлы) и пополнять ими библиотеку.

Спасибо за внимание!