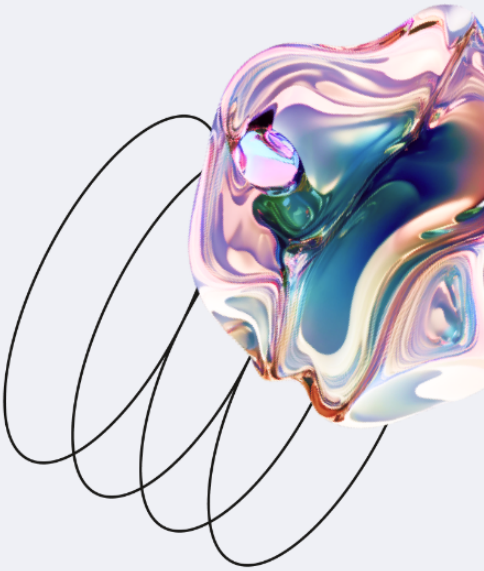
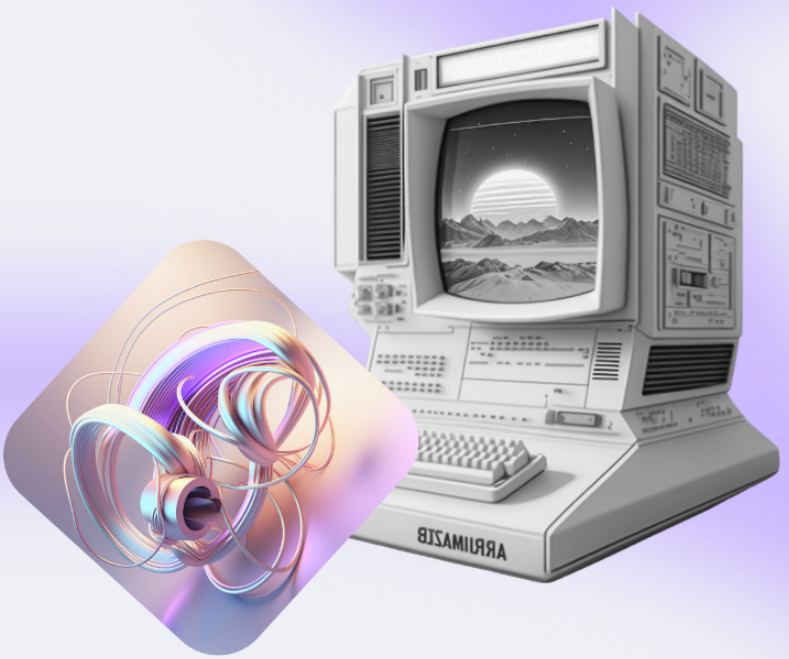
GeekBrains



Инструменты

BigData



Оглавление

Введение 3

Термины, используемые в лекции 3

Среда разработки для SQL 4

- JDBC-драйвер 5
- Hue 6
- Clouds 8

Среда разработки для Python 9

- Jupyter notebooks 9
- Apache Zeppelin 10
- Polynote 11
- JetBrains Big Data Tools 12

Визуализация 13

Домашнее задание 14

Используемая литература 14

Введение

Весь процесс работы с большими данными можно разделить на три этапа:

1. Сбор файлов разного формата из доступных источников.
2. Выбор способа хранения и размещение в хранилище.
3. Анализ и обработка результатов.

У необработанных данных нет ценности. Для преобразования и анализа Big Data нужны инструменты, придающие объёму информации полезную структуру.

Мы уже знакомы с основными фреймворками и инструментами, которые используются для сбора, хранения и обработки данных. Но для аналитики нам ещё нужны инструменты для EDA (разведочного анализа данных).

На этой лекции мы посмотрим на основные инструменты для EDA и визуализации:

- Jupyter notebooks,
- Apache Zeppelin,
- Superset,
- и прочие.

Термины, используемые в лекции

EDA (Exploratory Data Analysis) — разведочный анализ данных. Анализ основных свойств данных, нахождение в них общих закономерностей, распределений и аномалий, построение начальных моделей, зачастую с использованием инструментов визуализации.

CLI (Command line interface) — интерфейс командной строки. Способ взаимодействия между человеком и компьютером путём отправки компьютеру команд, представляющих собой последовательность символов. Команды интерпретируются с помощью специального интерпретатора — оболочки.

API (Application Programming Interface) — описание способов взаимодействия одной компьютерной программы с другими.

IDE (Integrated development environment) — единая среда разработки (ECP). Комплекс программных средств, которые программисты используют для разработки программного обеспечения.

BI (Business intelligence) — обозначение компьютерных методов и инструментов для организаций, обеспечивающих перевод транзакционной деловой информации в форму, которую удобно читать человеку. А также средства для массовой работы с такой обработанной информацией.

Среда разработки для SQL

В прошлой лекции мы узнали про SQL-инструменты:

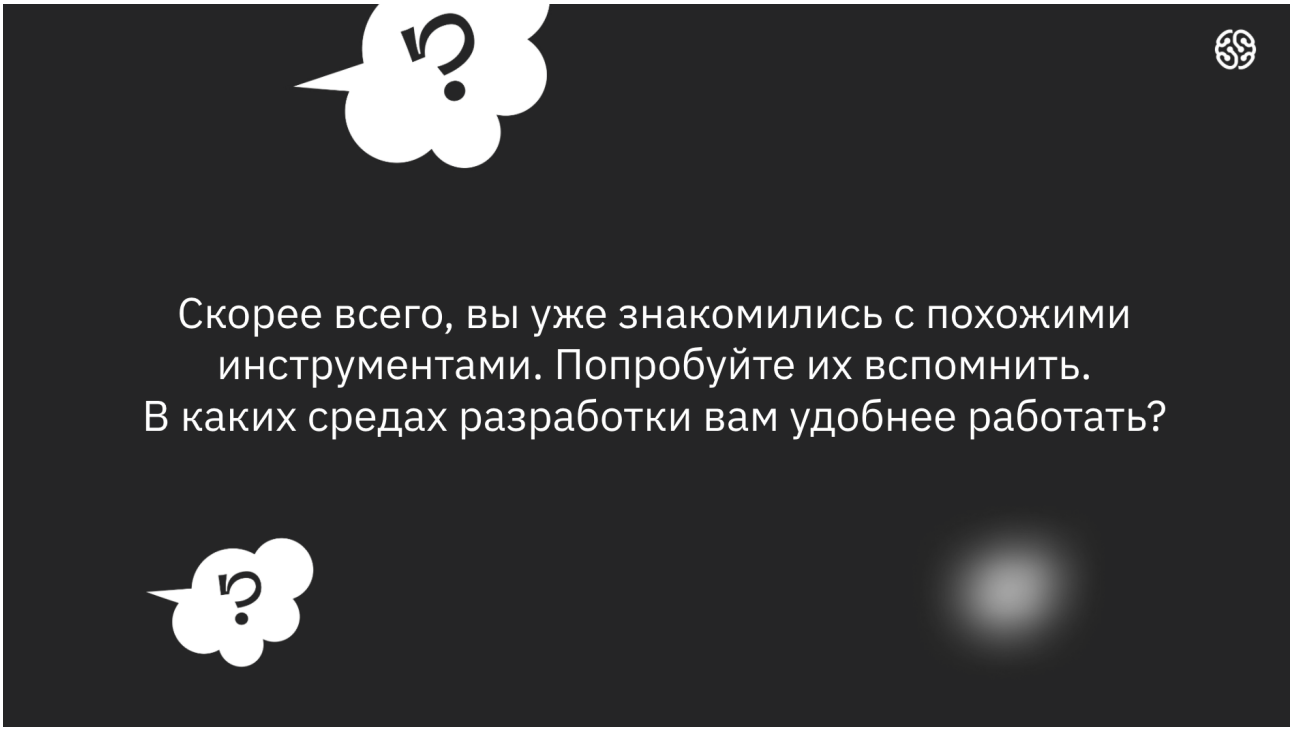
- Hive,
- Impala,
- Spark SQL,
- Cassandra,
- Drill,
- Presto.

Это всё SQL-движки (SQL engine), которые позволяют выполнять SQL или похожие на SQL запросы (query processor) на различных данных: например, SQL поверх CSV или HDFS. Некоторые движки также содержат в себе систему хранения данных (storage engine).

SQL-движки, за редким исключением, предоставляют только API и CLI. Эти интерфейсы подходят для написания скриптов, но не очень удобны для аналитики и изучения данных. Для бóльшего удобства были созданы среды разработки (IDE), которые, в свою очередь, можно поделить на две большие группы:

- **IDE для написания SQL-скриптов**. Они предоставляют графический редактор для скриптов (с подсказками, автодополнением и прочим) и удобное отображение результатов в виде таблиц, которые можно сортировать, фильтровать, редактировать и так далее.
- **IDE для языка программирования**, для которого есть API или коннектор к движку.

Основное отличие между этими группами в том, что во втором случае нам доступны все прочие инструменты для работы с данными, которые есть для языка программирования. А плюс SQL IDE в том, что нам не требуется знать ничего кроме самого SQL. Для простых задач SQL IDE вполне достаточно, но для сложной аналитики не всегда хватает функционала или он недостаточно удобен.



Некоторые языки программирования для анализа данных подходят лучше, некоторые хуже. В этой лекции мы будем использовать Python, но кроме Python есть и другие, например, Scala, Java или R.

> 💡 Сейчас Python — самый популярный язык для анализа данных. У него много библиотек для аналитики и визуализации: numpy, pandas, matplotlib, plotly и другие. Для Python практически всегда есть библиотеки для подключения к различным БД.

Сред разработки для SQL множество, но мы рассмотрим только те, которые чаще всего используются в экосистеме больших данных.

JDBC-драйвер

JDBC-драйвер (Java DataBase Connectivity) — это реализация стандартизированного интерфейса для работы с базой данных. Этот интерфейс позволяет подключить базу данных к IDE.

Стоит учитывать, что JDBC не создавался для больших данных. Если результат вашего запроса будет объёмным, IDE может просто не переварить такой ответ от БД.

> 💡 Hive имеет JDBC-драйвер. Его, например, можно подключить к следующим IDE:
>
> - Oracle SQL Developer,
> - DataGrip
> - и многим другим

Не все IDE используют JDBC-драйвер, у некоторых есть нативная поддержка. Чтобы узнать, поддерживает ли IDE нужный вам SQL-движок, посмотрите документацию IDE.

Hue

HUE (Hadoop User Experience) – это open source веб-интерфейс для анализа данных. Выпускается под лицензией Apache. Принадлежит компании Cloudera. Аббревиатуру HUE можно перевести как «Хадупный пользовательский опыт».

HUE — это пример браузерной IDE, которая была создана как раз для работы с большими данными в экосистеме Hadoop. В ней есть поддержка множества источников данных:

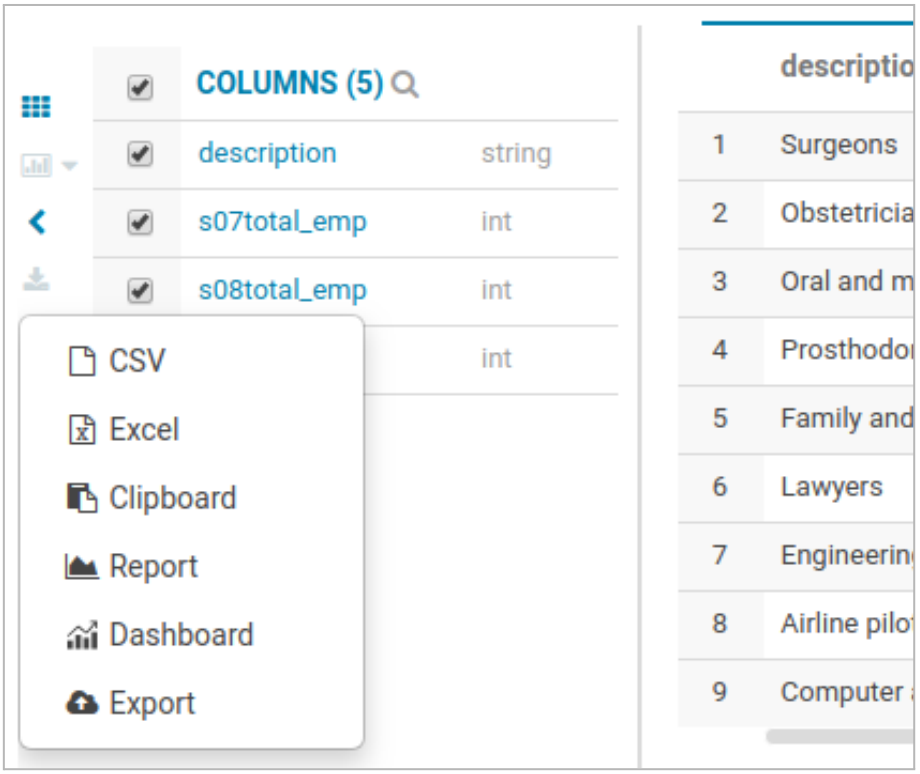
- Hive
- Impala
- Druid
- ksqlDB
- Flink SQL
- Spark SQL
- Phoenix SQL
- MySQL
- PostgreSQL
- Presto
- Dask-sql
- Elasticsearch SQL
- Calcite
- Athena
- Redshift
- Snowflake
- Big Query
- Oracle

💡 Описание, примеры и демо можно посмотреть на сайте gethue.com

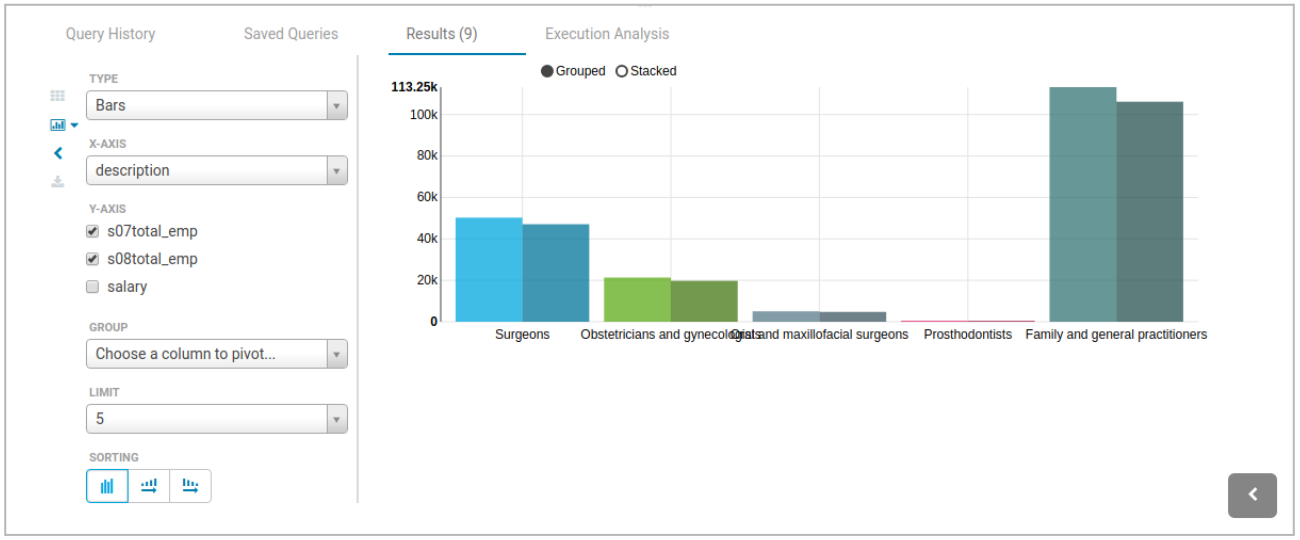
Функционал HUE:

- редактор SQL-запросов — проверка синтаксиса, справка, подсказки, автодополнение;

- работа с таблицей — просмотр, поиск, фильтрация, выгрузка;



- визуализация данных;



- создание пайплайнов и планировщик задач;
- работа с файлами в HDFS (файловый браузер);
- история запросов;
- мониторинг и статистика.

Задание: зайдите в демо-версию и изучите функционал HUE: user quick start guide.

Clouds

У облачных решений (Amazon DynamoDB, SnowFlake, Amazon Redshift, BigQuery) обычно есть своя браузерная / облачная IDE. Но также есть сторонние облачные IDE, куда можно подключить различные источники:

- Redash,
- Metabase,
- Popsql,
- Retool.

Обычно облачные IDE — смесь BI и IDE. Мы запомним, что такие существуют, но останавливаться подробно на них не будем, потому что они используются относительно редко.

Не всегда в браузере удобно работать. Примеры IDE, к которым можно подключить облачные хранилища:

- Aqua Data Studio,
- NoSQL Workbench,
- Dynobase.

💡 Облачные базы данных также имеют ODBC или JDBC-драйвера.

Среда разработки для Python

Как мы уже говорили выше, мы будем использовать Python и, соответственно, инструменты для Python.

Для анализа данных обычно используются ноутбуки, а не стандартные IDE для разработки. Ноутбук — это среда разработки, где сразу можно видеть результат выполнения кода и его фрагментов. Отличие от традиционной среды разработки в том, что код можно разбить на куски и выполнять их в произвольном порядке.

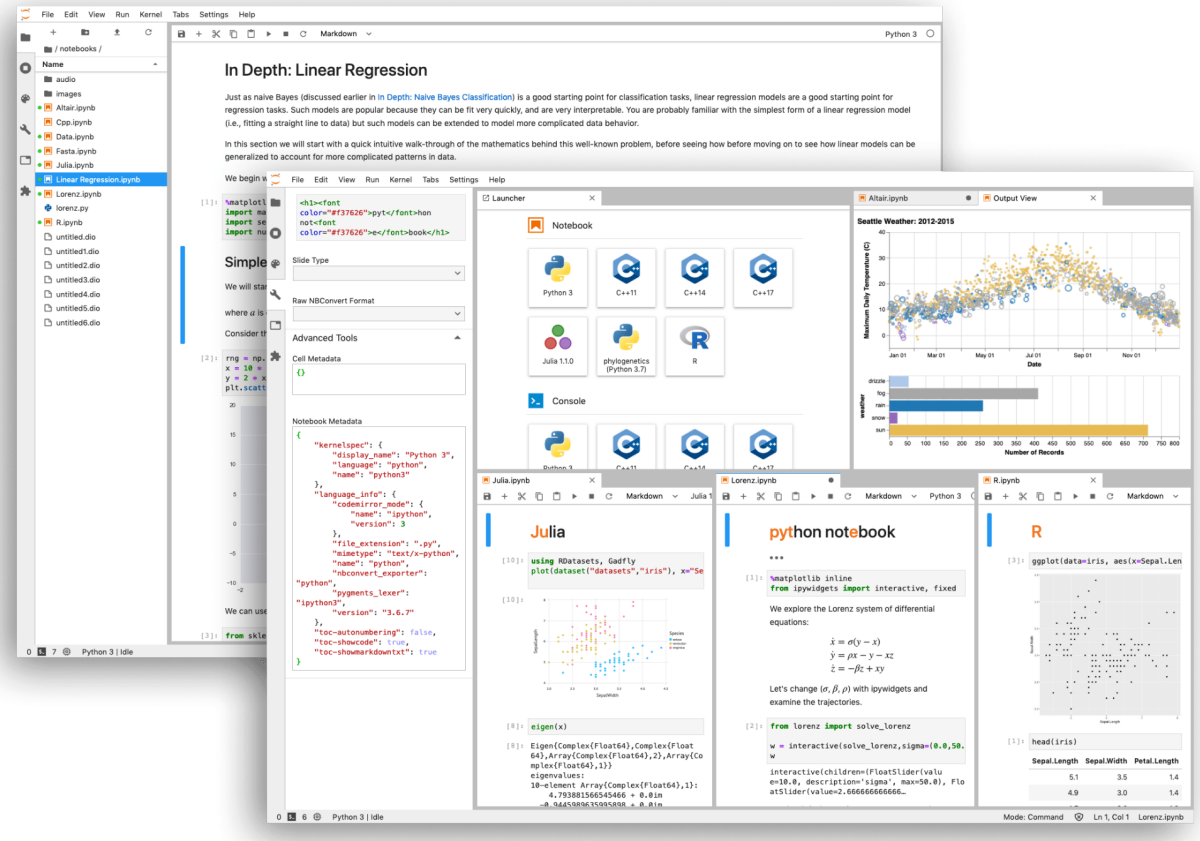
В такой среде разработки можно, например, написать функцию и сразу проверить её работу без запуска программы целиком. Или поменять порядок выполнения кода. Можно отдельно загрузить файл в память, отдельно проверить его содержимое, отдельно обработать содержимое. В ноутбуках также есть вывод результата сразу после фрагмента кода и форматирование комментариев. Например, можно прямо в середине кода увидеть построенный график, получить предварительные цифры или любую другую визуализацию.

💡 Перечисленные ниже инструменты поддерживают не только Python, но и Scala, R и другие языки

Jupyter notebooks

Jupyter — интерактивный блокнот, первоначально являвшийся веб-реализацией и развитием IPython, ставший самостоятельным проектом, ориентированным на работу со множеством сред выполнения — не только Python, но и R, Julia, Scala и ряда других.

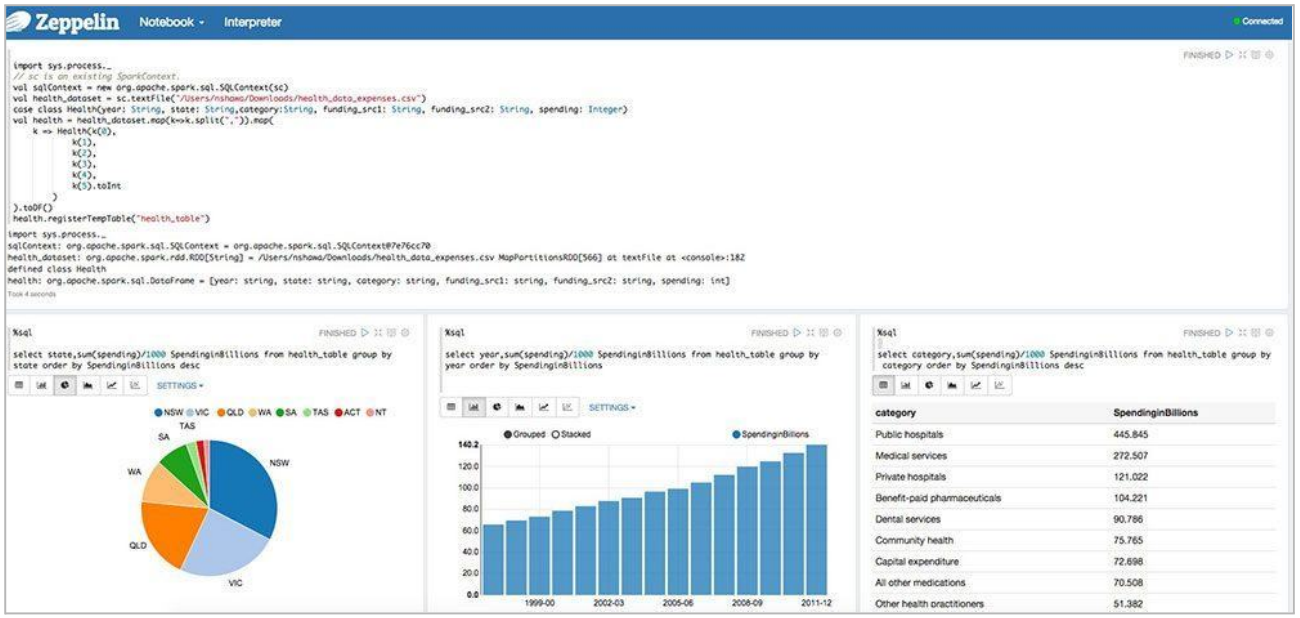
Чаще всего jupyter-ноутбуки применяют в машинном обучении, например, при подготовке нейросетей. Ещё их используют специалисты по data science и начинающие программисты на Python. Сильная сторона этой среды разработки — визуализация данных.



Apache Zeppelin

Apache Zeppelin — интерактивный блокнот для анализа и визуализации данных, а также совместной работы над данными с использованием средств Apache Spark. Позиционируется как аналог Jupyter для экосистемы Hadoop.

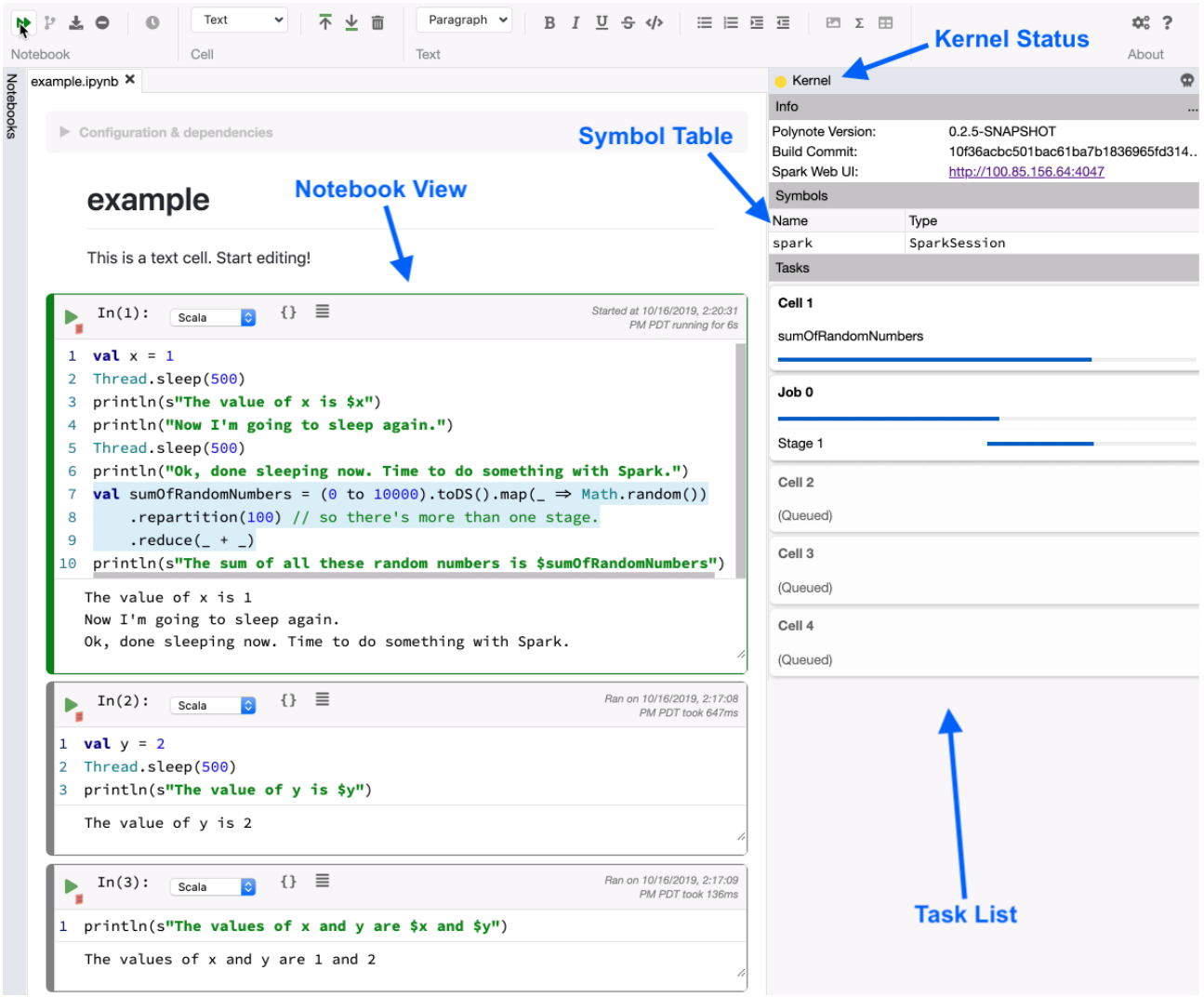
Комбинирование различных источников данных в рамках одного дашборда — одно из его ключевых преимуществ. Большинство компаний с уже работающими БД и системами аналитики могут использовать его «из коробки». Энтузиасты с более экзотичными БД могут написать интерпретатор самостоятельно, о чём на сайте продукта есть статья.



Polynote

Polynote — это интерактивная среда вычислений с блочной структурой. В каждом блоке может быть либо текст, либо код на каком-либо языке программирования. Содержимое в каждом из блоков выполняется по отдельности. Блоки можно редактировать, удалять, добавлять новые или менять порядок. При этом от результата вычислений в одном блоке зависит результат в последующих.

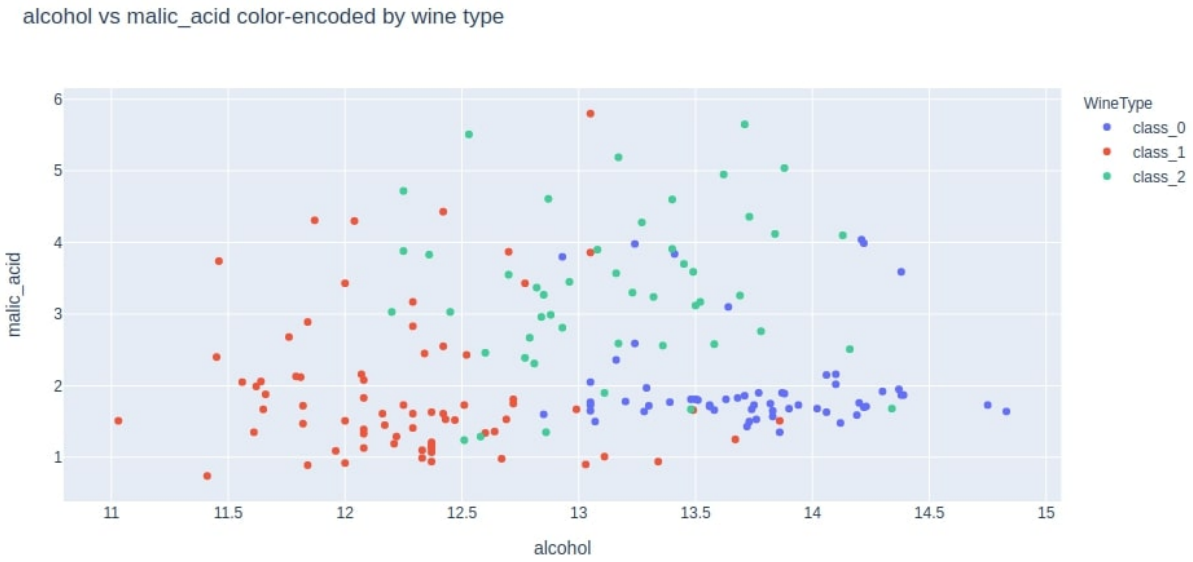
В Polynote есть много функций для работы и с кодом, и с текстом. Инструмент можно воспринимать как гибрид IDE и текстового редактора. У него есть автодополнение кода, подсветка ошибок, подсказки для параметров функций и методов.



JetBrains Big Data Tools

Big Data Tools — это плагин, позволяющий подключаться к кластерам Hadoop и Spark. Предоставляет возможность мониторинга узлов, приложений и отдельных задач. Кроме того, можно создавать, запускать и редактировать ноутбуки Zeppelin. Можно не переключаться на веб-интерфейс Zeppelin, а продолжать спокойно работать из любимой IDE. Плагин обеспечивает удобную навигацию по коду, умное автодополнение, рефакторинги и квик-фиксы прямо внутри ноутбука.

Визуализация



Источник: How to Create Basic Dashboard in Python with Widgets [plotly & Dash]?

Для анализа данных требуется визуализация. График может быть нагляднее цифр.

При разработке в Python можно использовать библиотеки:

- matplotlib
- plotly
- dash
- seaborn

Также можно использовать BI-инструменты (примеры open source) при наличии соответствующего коннектора для вашей БД:

- Apache Superset
- Preset
- Metabase
- Redash
- Tipboard

Домашнее задание

Выберите хотя бы по одному инструменту для SQL и для Python. Изучите документацию.

Используемая литература

1. Hue
2. Jupyter Notebook
3. Apache Zeppelin
4. Polynote
5. Big Data Tools - IntelliJ IDEs Plugin | Marketplace
6. Apache Hive Essentials [Book]
7. DataGrip: The Cross-Platform IDE for Databases & SQL by JetBrains
8. Обзор плагина Big Data Tools