Matplotlib, Seaborn, SQL - введение

Преподаватель: Герард Костин

Matplotlib

Matplotlib - потрясающая библиотека визуализации на Python для 2D-графиков массивов.

• Matplotlib - это многоплатформенная библиотека визуализации данных, построенная на массивах NumPy и предназначенная для работы с более широким стеком SciPy.

• Одним из самых больших преимуществ визуализации является то, что она позволяет нам получить визуальный доступ к огромным объемам данных в легко усваиваемых визуальных эффектах.

• Matplotlib состоит из нескольких типов графиков, таких как line, plot, scatter,bar, histogram и т. Д.

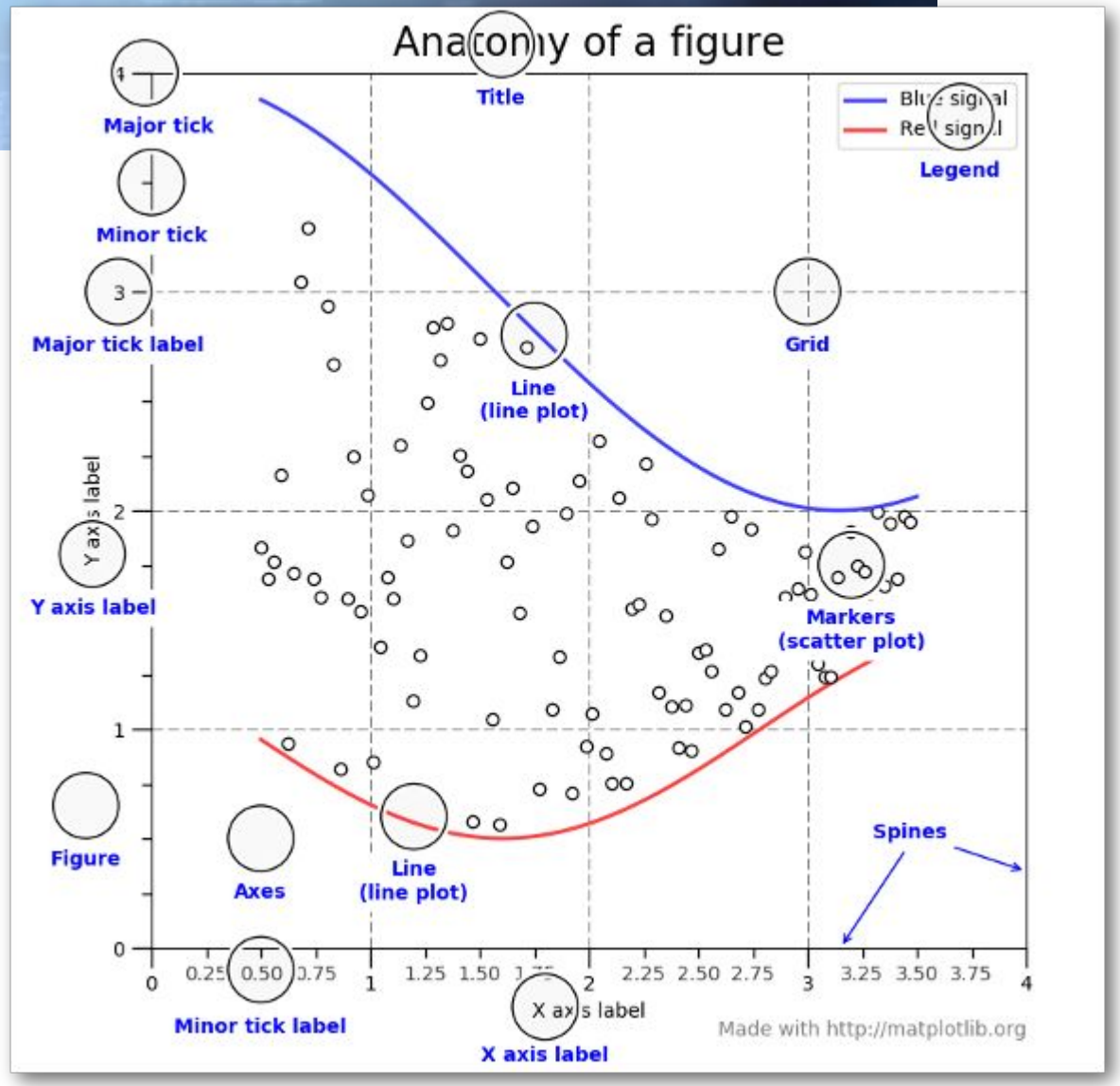
Matplotlib - бесплатный!

Matplotlib строит кривые, очень похожие на MATLAB.

Единственное отличие - это Matrix Laboratory, или MATLAB требует лицензии и стоит очень дорого.

Исходный код распространяется под лицензией BSD.

Терминология



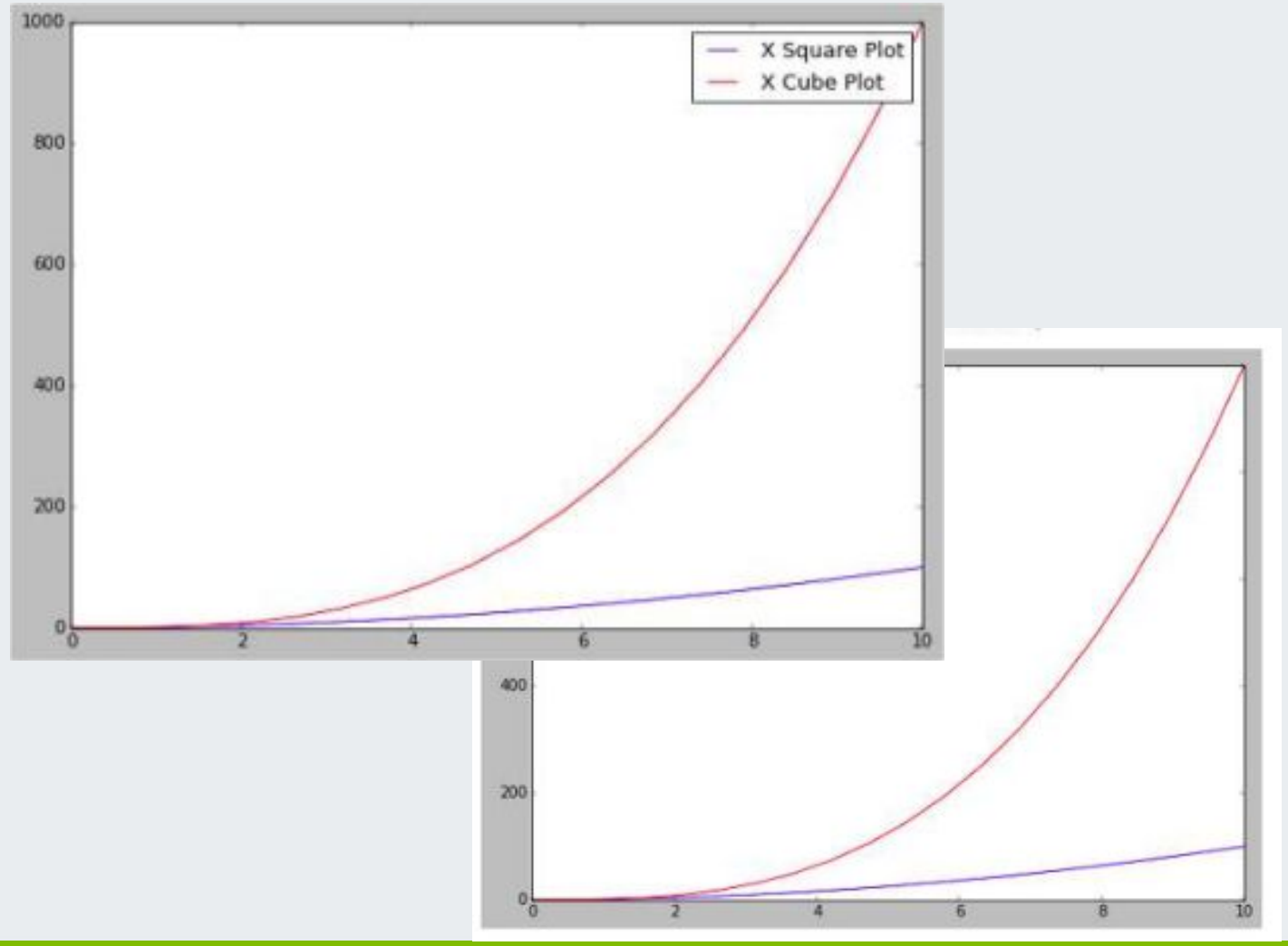
- Figure - это окончательное изображение, которое может содержать 1 или более *осей* (*axes*).
- Axes (оси) представляют собой отдельный график (*plot*).

Легенда

Легенды позволяют различать графики между собой.

С помощью Legends вы можете использовать тексты меток, чтобы идентифицировать или отличать один график от другого.

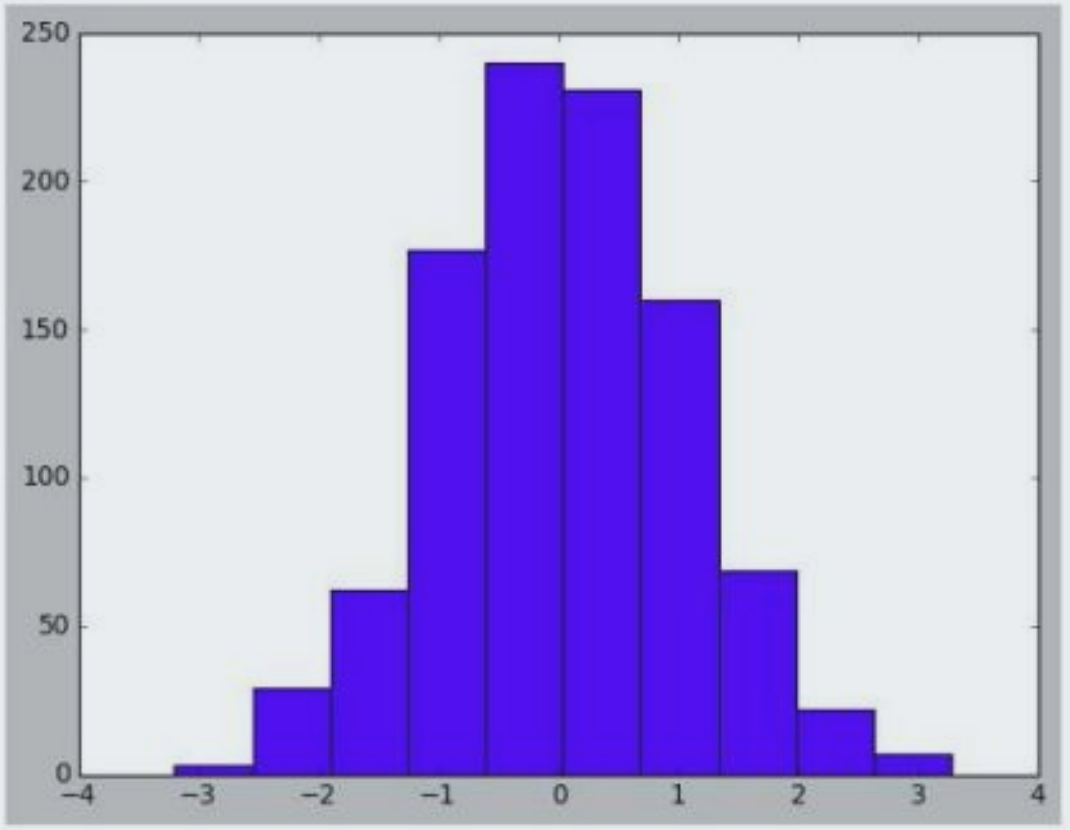
Например, предположим, что у нас есть фигура, состоящая из двух графиков, как показано ниже:



Histogram

- **ГИСТОГРАММА ПОМОГАЕТ НАМ ПОНИМАТЬ РАСПРЕДЕЛЕНИЕ ЧИСЛОВОГО ЗНАЧЕНИЯ СПОСОБОМ, КОТОРЫМ ВЫ НЕ МОЖЕТЕ ДЕЛАТЬ, ИСПОЛЬЗУЯ СРЕДНЕЕ, МЕДИАНУ или МОДУ**

```
x = np.random.randn(1000)
plt.hist(x);
```



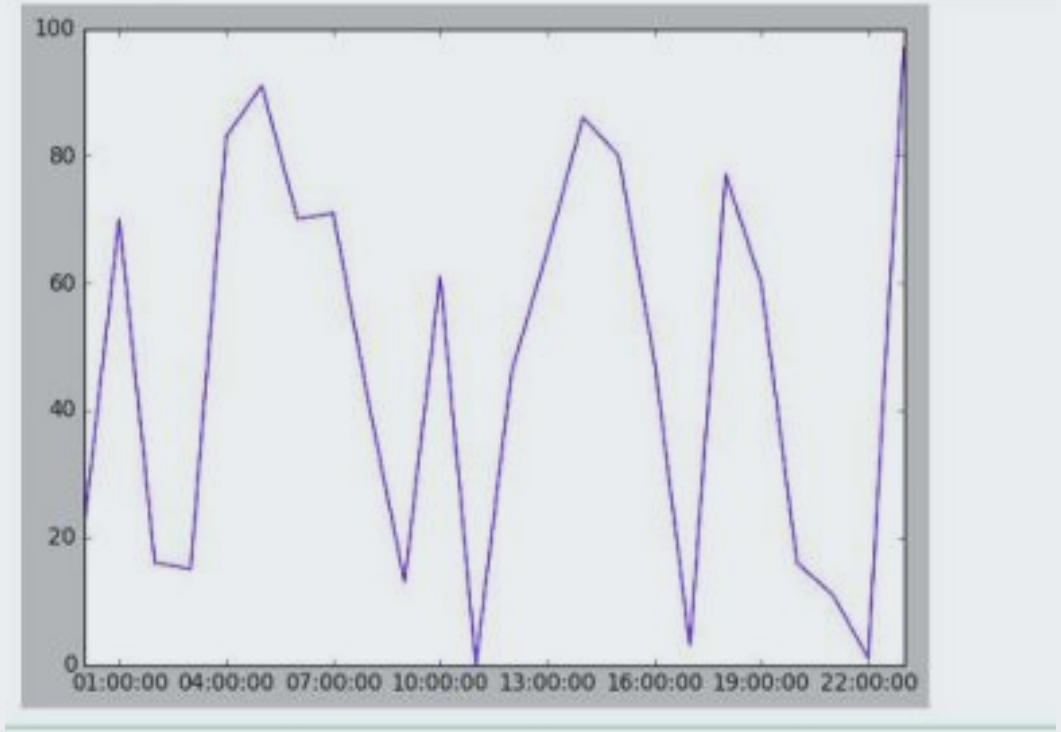
Временной ряд

- представляет собой диаграмму, показывающую тенденцию за определенный период времени.
- Он позволяет вам проверять различные гипотезы при определенных условиях, например, что происходит в разные дни недели или в разное время дня.

```
import matplotlib.pyplot as plt
import datetime
import numpy as np

x = np.array([datetime.datetime(2018, 9, 28, i, 0) for i in range(24)])
y = np.random.randint(100, size=x.shape)

plt.plot(x,y)
plt.show()
```

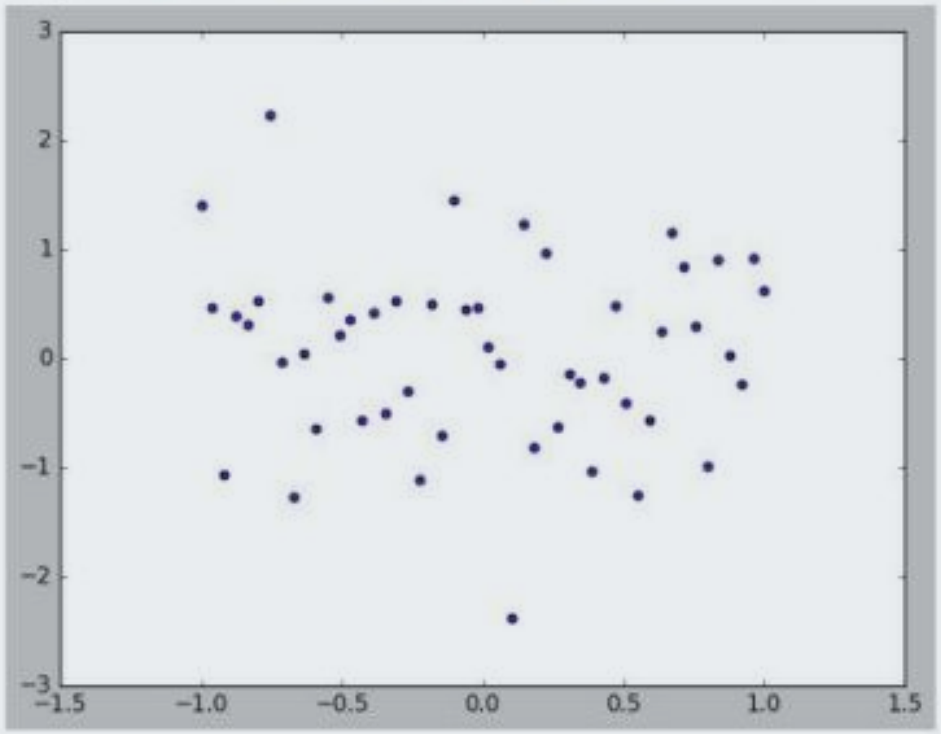


Scatter plot

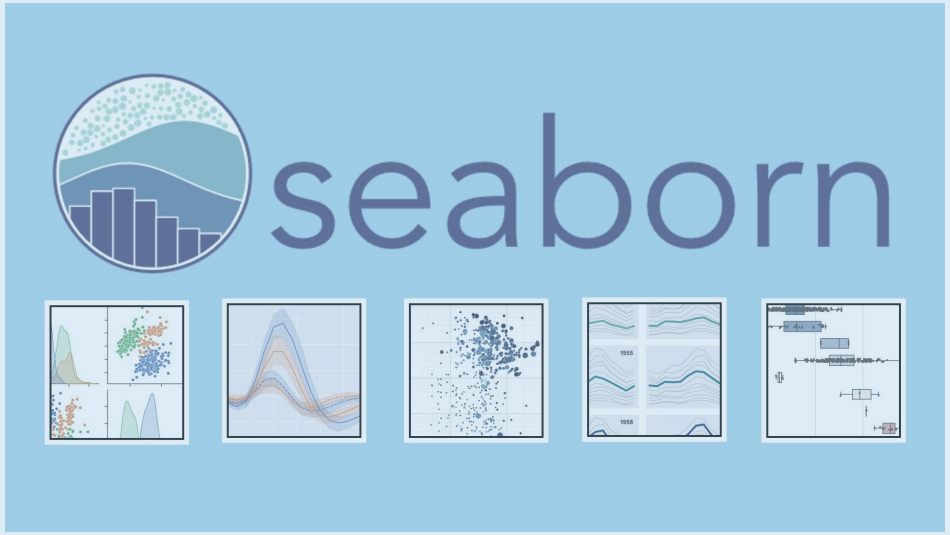
- предлагает удобный способ визуализировать, как два числовых значения связаны в ваших данных.
- Это помогает понять отношения между несколькими переменными.
- Используя метод .scatter (), мы можем создать диаграмму.

```
fig, ax = plt.subplots()
x = np.linspace(-1, 1, 50)
y = np.random.randn(50)
ax.scatter(x, y)
```

```
<matplotlib.collections.PathCollection at 0x7f004eaa30f0>
```



Matplotlib vs Seaborn

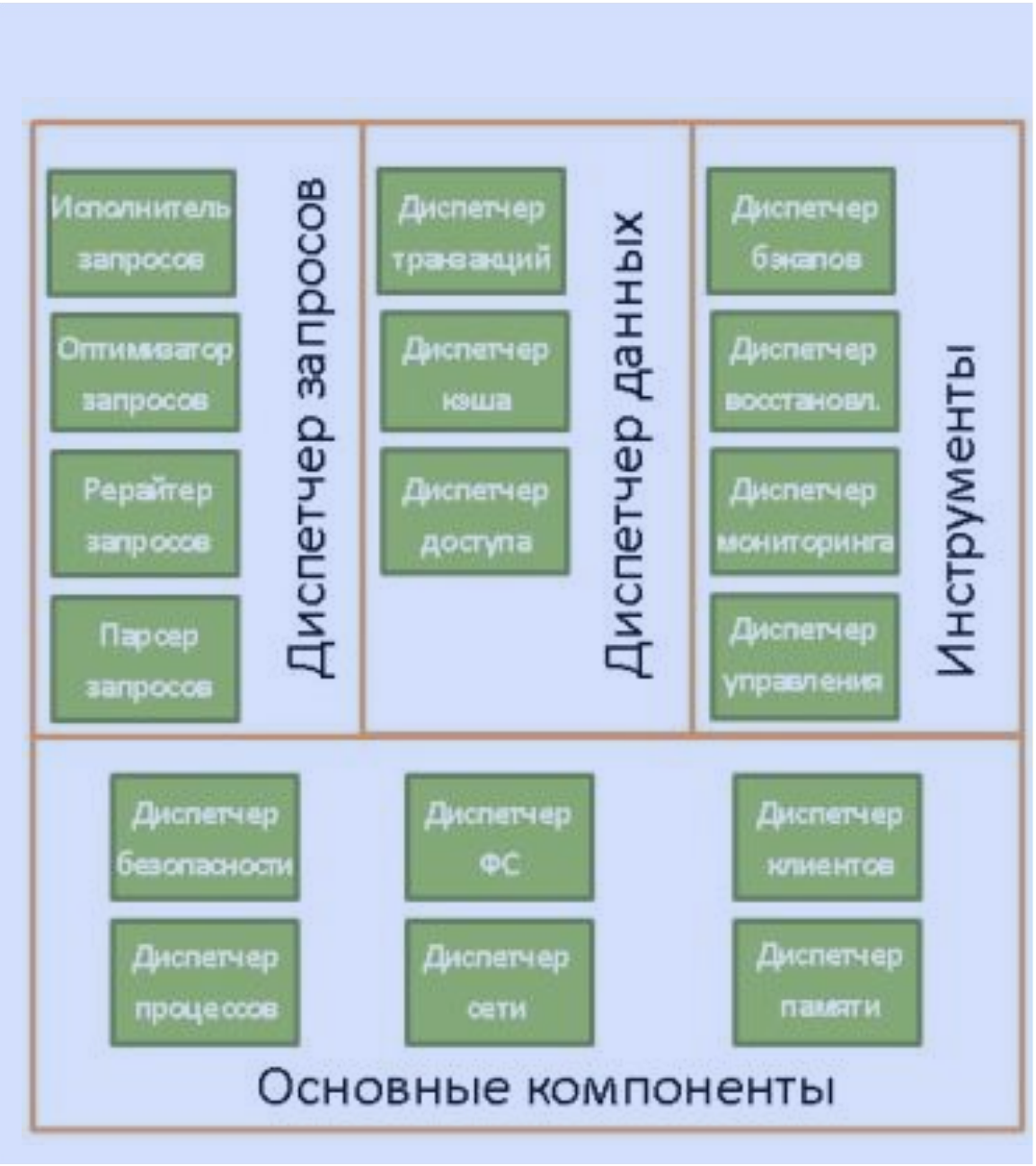


Seaborn использует интересные темы и стили, а matplotlib используется для создания базовых графиков.

Seaborn содержит несколько готовых графиков и шаблонов для визуализации данных, тогда как в matplotlib наборы данных визуализируются с помощью линий, точечных диаграмм, круговых диаграмм, гистограмм, гистограмм и т. Д.

Seaborn более удобен в Pandas

Seaborn значительно более организован и функционален, чем Matplotlib, и обрабатывает весь набор данных как отдельную единицу.



Реляционные БД

БД (База Данных) представляет собой совокупность информации, к которой можно легко получить доступ и модифицировать.

Особенности:

- Использование **транзакции** для обеспечения сохранности и связанности данных.
- Быстрая обработка данных вне зависимости от их объёма.

ОСНОВНЫЕ КОМПОНЕНТЫ БД

- **Диспетчер процессов**. Во многих БД имеется пул процессов/потоков, которыми нужно управлять. Причём в погоне за производительностью некоторые БД используют свои собственные потоки, а не предоставляемые ОС.
- **Диспетчер сети**. Пропускная способность сети имеет большое значение, особенно для распределённых БД.
- **Диспетчер файловой системы**. Первым «бутылочным горлышком» любой БД является производительность дисковой подсистемы. Поэтому очень важно иметь диспетчер, который идеально работает с файловой системой ОС или даже заменяет её.
- **Диспетчер памяти**. Чтобы не упереться в невысокую производительность дисковой подсистемы, нужно иметь много оперативной памяти. А значит, нужно эффективно ею управлять, что и делает данный диспетчер. Особенно когда у вас много одновременных запросов, использующих память.
- **Диспетчер безопасности**. Управляет аутентификацией и авторизацией пользователей.
- **Диспетчер клиентов**. Управляет клиентскими соединениями.

СВОЙСТВА ТРАНЗАКЦИЙ

ACID-транзакция (Atomicity, Isolation, Durability, Consistency) — это элементарная операция, единица работы, которая удовлетворяет 4 условиям:

- **Атомарность (Atomicity).** Нет ничего «меньше» транзакции, никакой более мелкой операции. Даже если транзакция длится 10 часов. В случае неудачного выполнения транзакции система возвращается в состояние «до», то есть транзакция откатывается.
- **Изолированность (Isolation)**. Если в одно время выполняются две транзакции A и B, то их результат не должен зависеть от того, завершилась ли одна из них до, во время или после исполнения другой.
- **Надёжность (Durability).** Когда транзакция зафиксирована (commited), то есть успешно завершена, использовавшиеся ею данные остаются в БД вне зависимости от возможных происшествий (ошибки, падения).
- **Консистентность (согласованность) (Consistency).** В БД записываются только валидные данные (с точки зрения реляционных и функциональных связей). Консистентность зависит от атомарности и изолированности.

Уровни работы с данными

Слой доступа к данным, который удобно использовать из языков программирования;

Слой хранения. Это отдельный слой, потому что обычно хранить данные удобно другими способами, чем использовать: эффективно по памяти, выравнивать, складывать на диск. Это к вопросу о schemaless: схема, которая удобна для хранения, не удобна для доступа.

«Железо» — слой, где лежат данные, причем там они организованы еще третьим способом, потому что дисками управляет операционная система, и общаются они только через драйвер. В этот уровень мы не будем сильно вникать.

Основы реляционной алгебры

Реляционной базой данных называется совокупность отношений, содержащих всю информацию, которая должна хранится в базе

Любая Таблица состоит из N строк, строка в таблице является кортежем в реляционной теории. Множество упорядоченных кортежей называется отношением.

Первичный ключ это атрибут или набор из минимального числа атрибутов, который однозначно идентифицирует конкретный кортеж и не содержит дополнительных атрибутов.

В реляционной БД таблицы взаимосвязаны и соотносятся друг с другом как главные и подчиненные.

Связь главной и подчиненнной таблицы осуществляется через первичный ключ (primary key) главной таблицы и внешний ключ (foreign key) подчиненной таблицы.

Внешний ключ это атрибут или набор атрибутов, который в главной таблице является первичным ключем.

Операции реляционной алгебры

- Выборка
- Проекция
- Декартово произведение
- Соединение
- Пересечение
- Вычитание
- и др.