Работа с файлами

Константин Башевой
Аналитик-разработчик, Яндекс

Константин Башевой
Аналитик-разработчик
Яндекс

Помогаю аналитикам с инфраструктурой

Собираю инструменты обработки данных

Рассказываю как это весело

Последние 10 лет:

- Rambler&Co
- Ростелеком
- Яндекс

Что сегодня будет

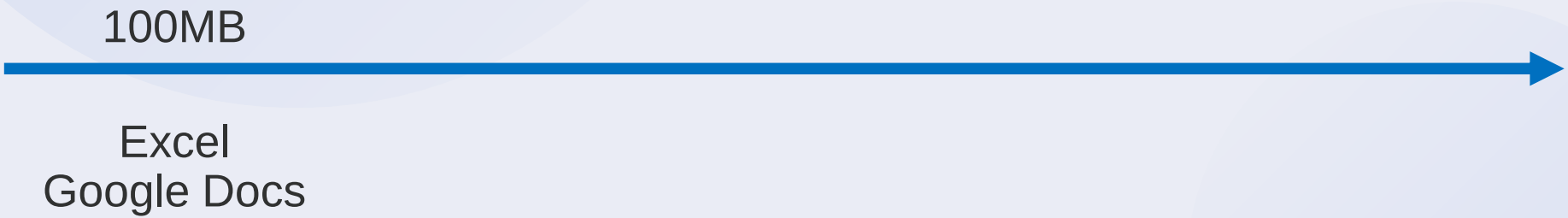
Программа на сегодня

Чтение файлов любого размера

Запись произвольных объектов в файл

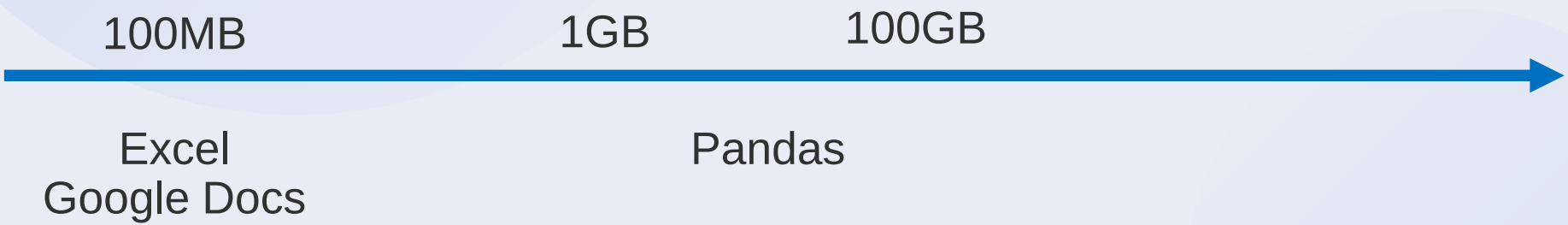
Пакетный менеджер pip

В чем вообще проблема файлы открывать?



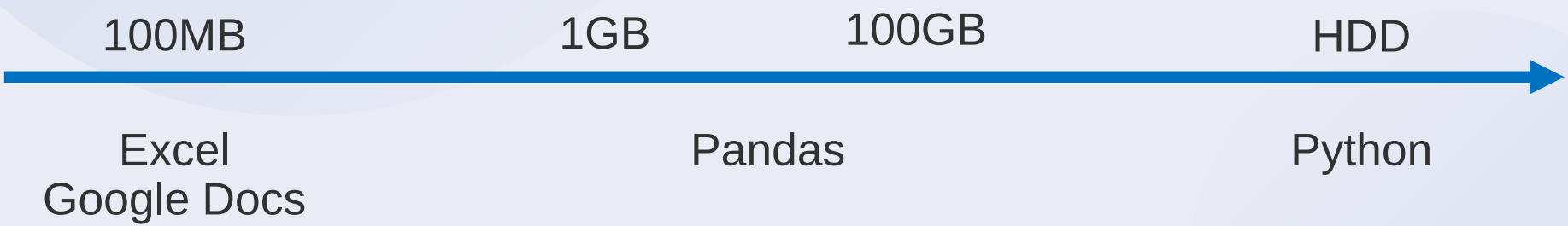
В чем вообще проблема файлы открывать?

Библиотека Pandas ограничена объемом RAM



В чем вообще проблема файлы открывать?

Библиотека Pandas ограничена объемом RAM



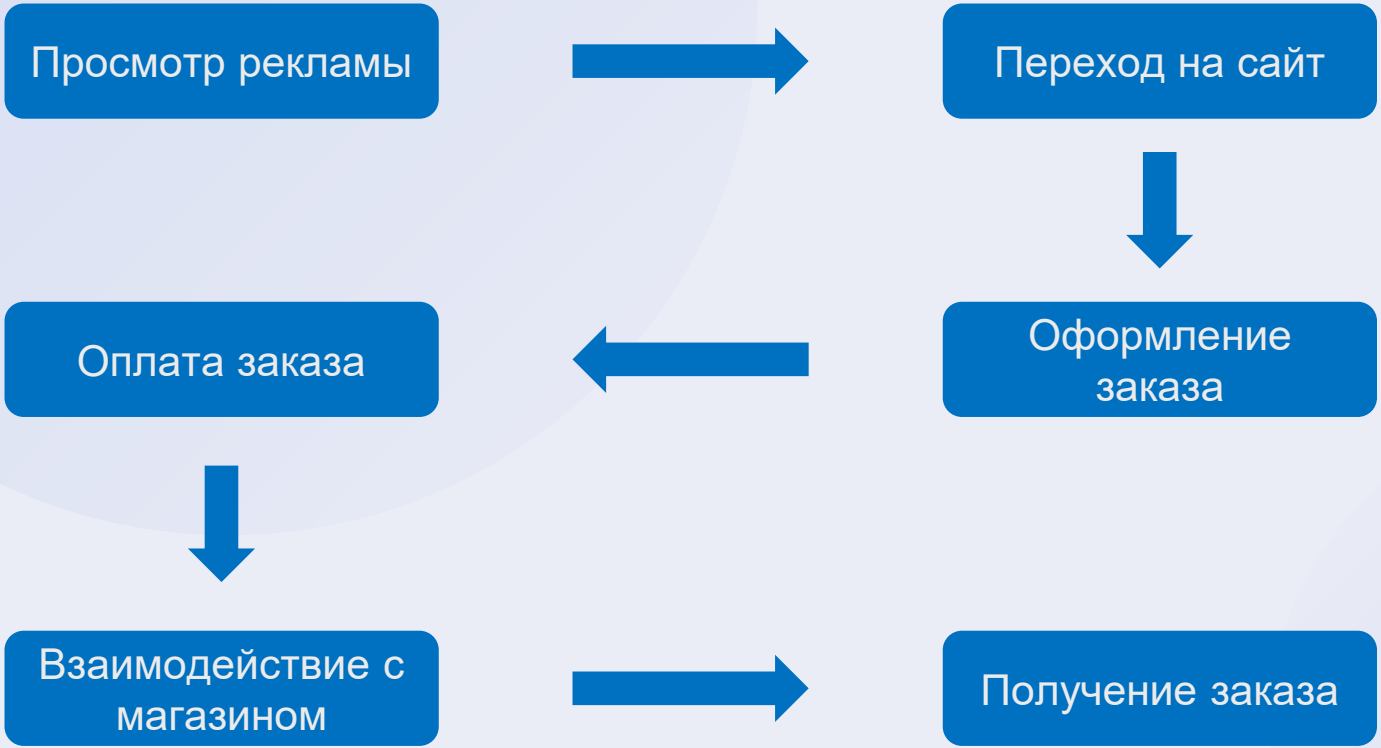
Постановка задачи

Сквозная аналитика

Информация о действиях пользователя на проекте собирается в разных системах.

Сквозная аналитика – объединение данных из разных систем в одну таблицу (в идеале).

Сквозная аналитика



Сквозная аналитика

Просмотр рекламы	Я.Директ MyTarget AdWords ВКонтакте
Переход на сайт	Я.Метрика Appmetrica G.Analytics
Оплата заказа	Биллинги 1С, Axapta Аудит
Оформление заказа	Каталог Рекомендации Склады Поставщики
Взаимодействие с магазином	Сторонние CRM и БД
Получение заказа	Курьеры Отказы

Зачем нужна одна таблица?

Данные в одной таблице удобны для решения многих задач.

Дата	User ID	Визиты	Покупки	Расходы	Источник	Город
2019-12-01	u1@yandex.ru	4	1	350	AdWords	Москва
2019-12-29	u2@mail.ru	15	2	2300	Я.Директ	Тула

Зачем нужна одна таблица?

Данные в одной таблице удобны для решения многих задач.

Дата	User ID	Визиты	Покупки	Расходы	Источник	Город
2019-12-01	u1@yandex.ru	4	1	350	AdWords	Москва
2019-12-29	u2@mail.ru	15	2	2300	Я.Директ	Тула

- Какой рекламный источник наиболее прибылен?
- В каком городе аномальная доля возвратов товара?
- Какие пользователи похожи на перекупщиков?
- Какого ассортимента не хватает в Краснодаре?

Исходные данные

Таблица визитов

User ID	Источник визита
e1bd168161	context
f697206af5	other

Таблица покупок

User ID	Категория покупки
e1bd168161	Электроника
dd3d43c266	Продукты

Проблемы

- Лог визитов очень большой (например, 100Гб). В RAM не влезет
- Лог покупок содержит словари

  {'user_id': '1840e0b9d4', 'category': 'Продукты'}

Общие рекомендации

Если позволяет задача читайте файл построчно

Осторожно используйте параметр mode=“w”. Файл чистится

Не пишите одновременно в один и тот же файл два потока