



УНИВЕРСИТЕТ
искусственного интеллекта

Методы получения данных из систем источников

План

- ETL/ELT общий подход
- Протокольный метод
- Поточный метод
- Метод запросов к данным
- APIs



ETL/ELT

ОБЩИЙ ПОДХОД

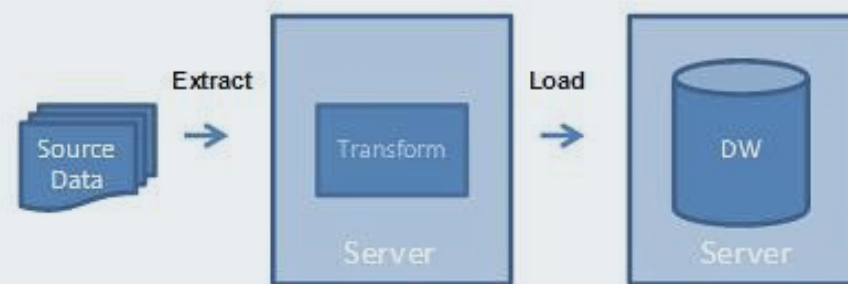
Две основные парадигмы
извлечения данных:

ETL – Extract Transform Load

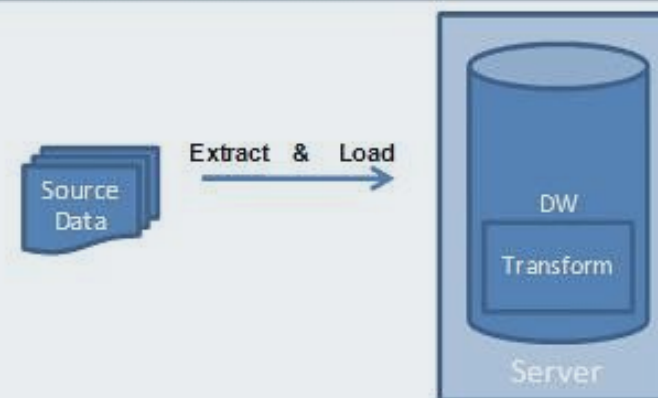
ELT – Extract Load Transform

- Data Engineer – ключевая роль

ETL vs. ELT as a Data Integration Approach

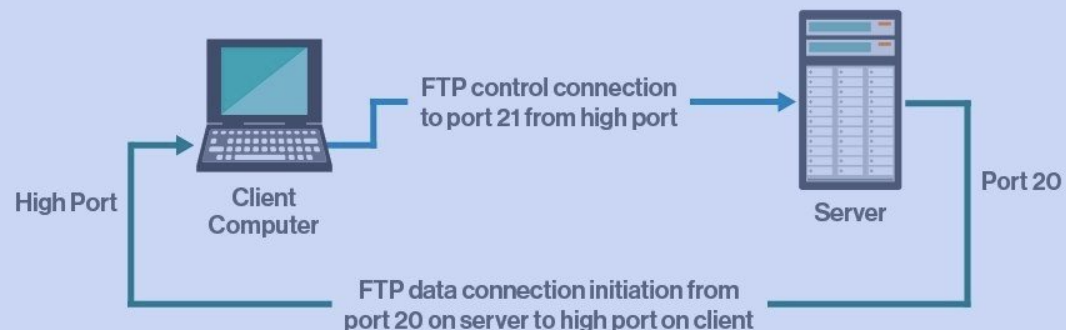


ETL – Transform on a separate H/W, S/W platform

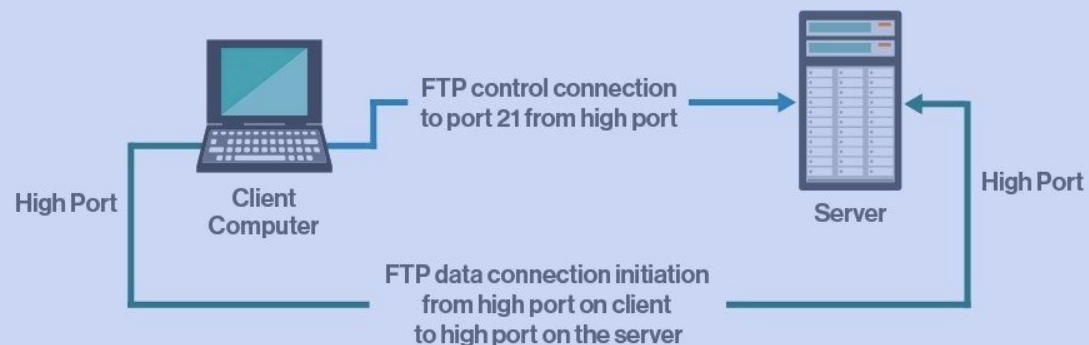


ELT – Load and Transform on the same H/W, S/W platform as the DW

Active FTP



Passive FTP



DESIGN: TECHTARGET/CHRISTOPHER SEERO

ПРОТОКОЛЬНЫЙ МЕТОД. **FTP**

FTP-сервер – сервер, работающий по File Transfer Protocol (протоколу передачи файлов).

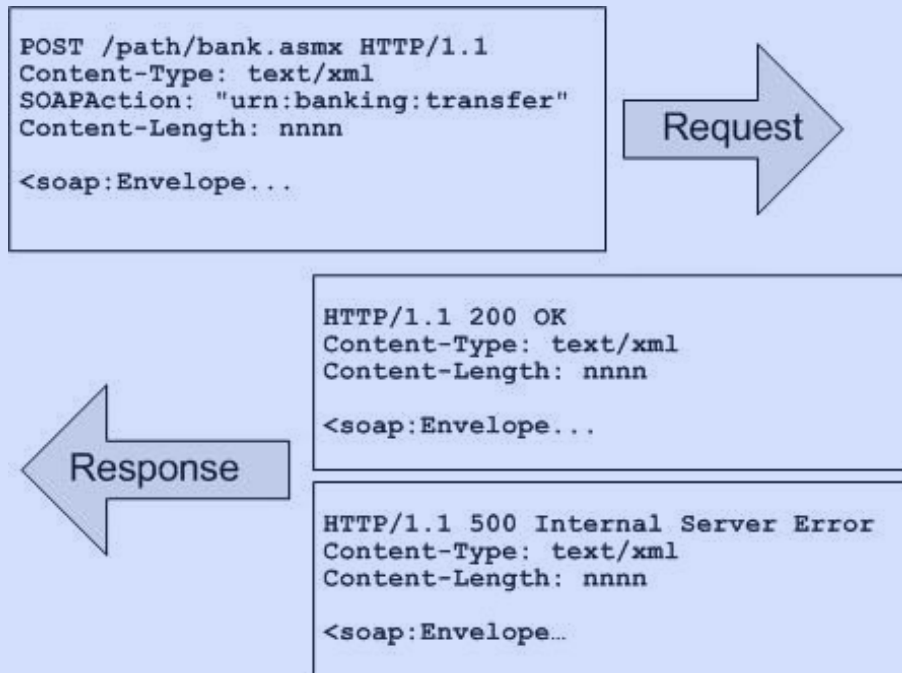
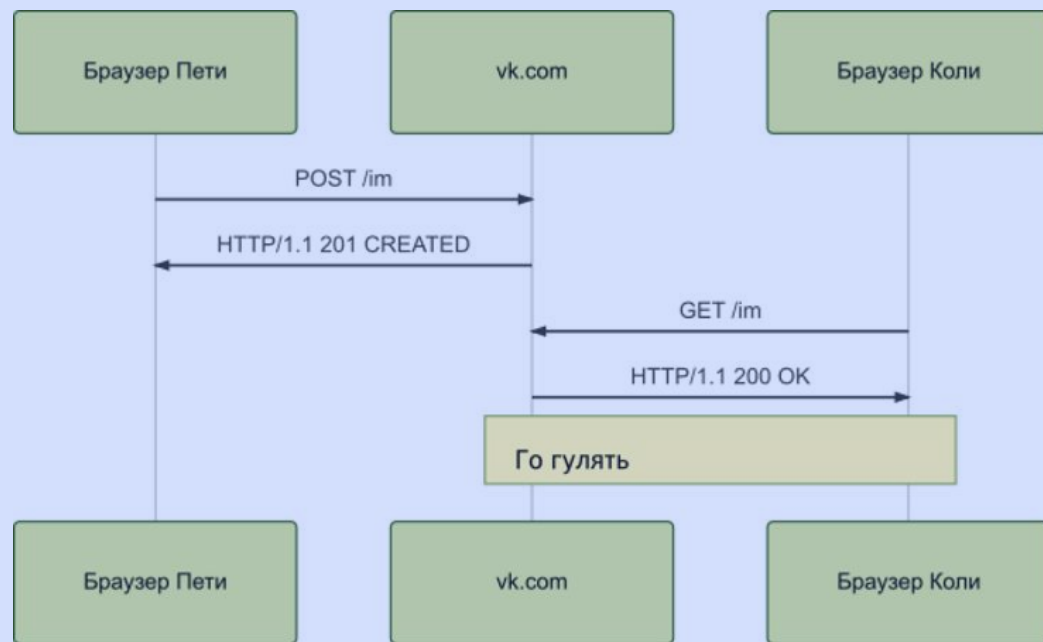
Используется для обмена файлами между компьютерами в сети. Работает по «клиент-сервер» модели.

Недостатки:

- Слабая защиты от взлома и атак.
- отсутствие понятия кодировки
- отсутствие встроенных механизмов шифрования
- Двухпортовость (проблемы с сетевыми экранами)

Плюсы:

- широкая распространённость,
- разнообразие серверов и клиентов
- меньшие накладные расходы при загрузке больших файлов



ПРОТОКОЛЬНЫЙ МЕТОД. HTTP

HTTP (HyperText Transport Protocol).

Протокол передачи данных, используемый обычно для получения информации с веб-сайтов.

HTTPS (от англ. **H**yper**T**ext **T**ransfer **P**rotocol **S**ecure) — расширение протокола HTTP, поддерживающее шифрование посредством (SSL и TLS).

Основные методы:

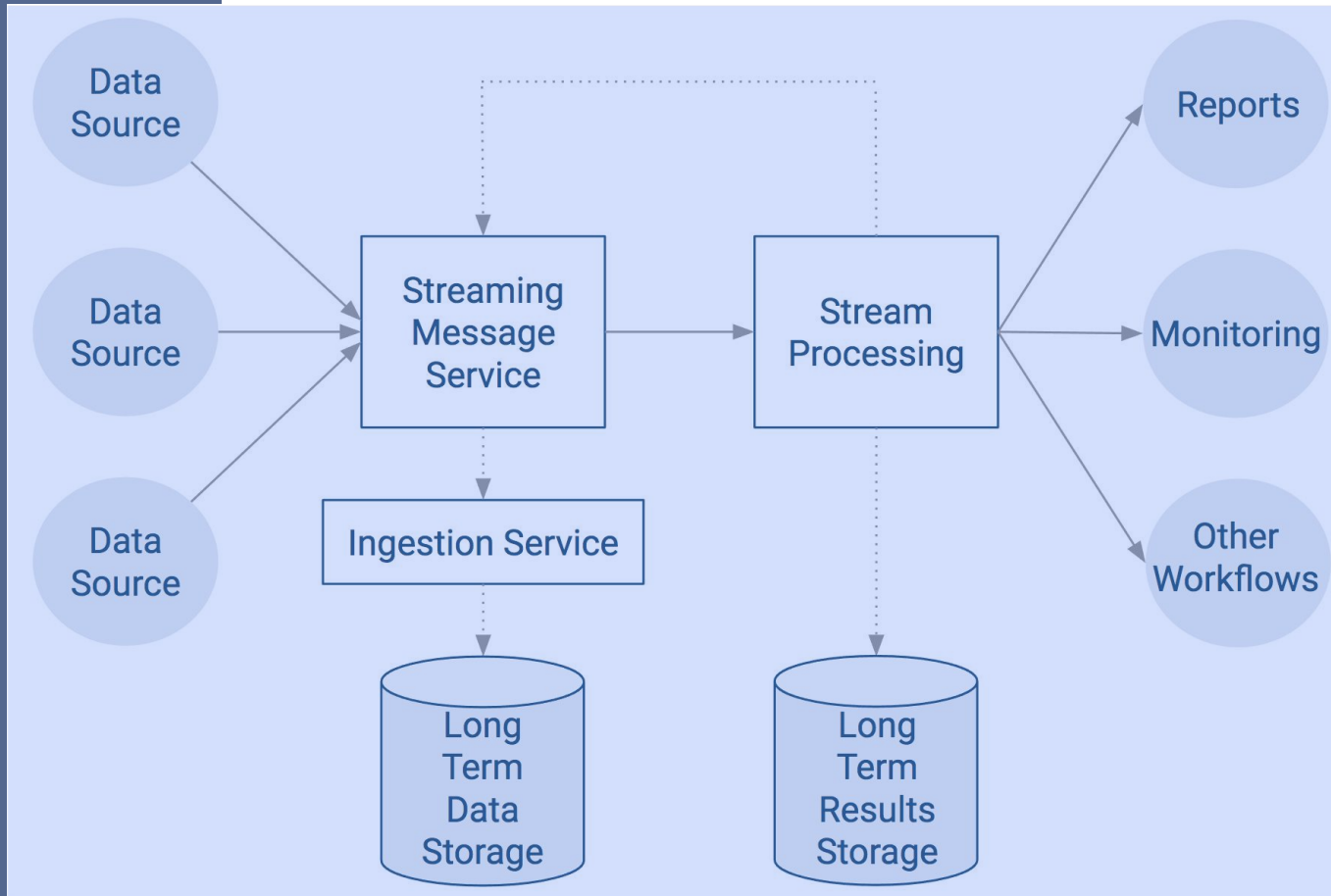
GET — запросы данных с определенного ресурса, на котором данные не изменяются.

POST — используется для отправки данных на сервер для создания ресурса.

PUT — метод для обновления существующего на сервере ресурса, используя содержимое тела запроса.

DELETE — удаляет определённый ресурс

Поточный метод получения данных



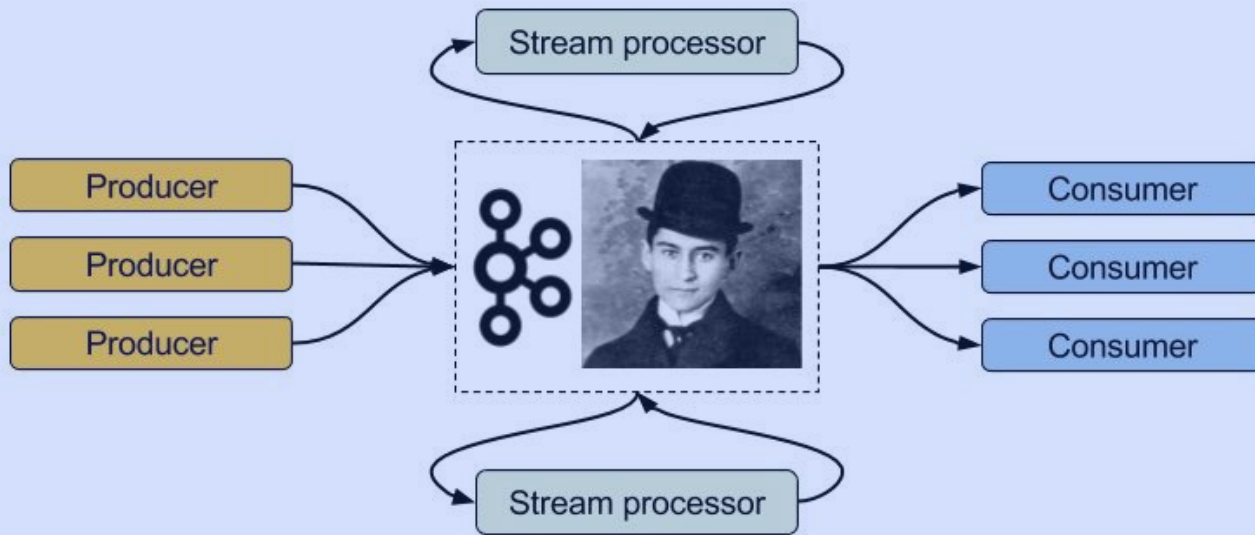
Особенности:

- Большое количество источников
- Появление события – простой пуассоновский процесс
- Одни и те же данные отдаются большому количеству потребителей
- Гарантированная доставка событий
- Доставка по факту возникновения события

ПОТОЧНЫЙ МЕТОД ПОЛУЧЕНИЯ ДАННЫХ

Apache KAFKA

Apache Kafka – диспетчер сообщений на Java платформе (распределенный журнал событий).



Основные особенности:

Топик (Topic) – набор партиций состоящий из журнал событий (commit log)

Партиция (Partition) – упорядоченный (сохраняется порядок сообщений) журнал событий

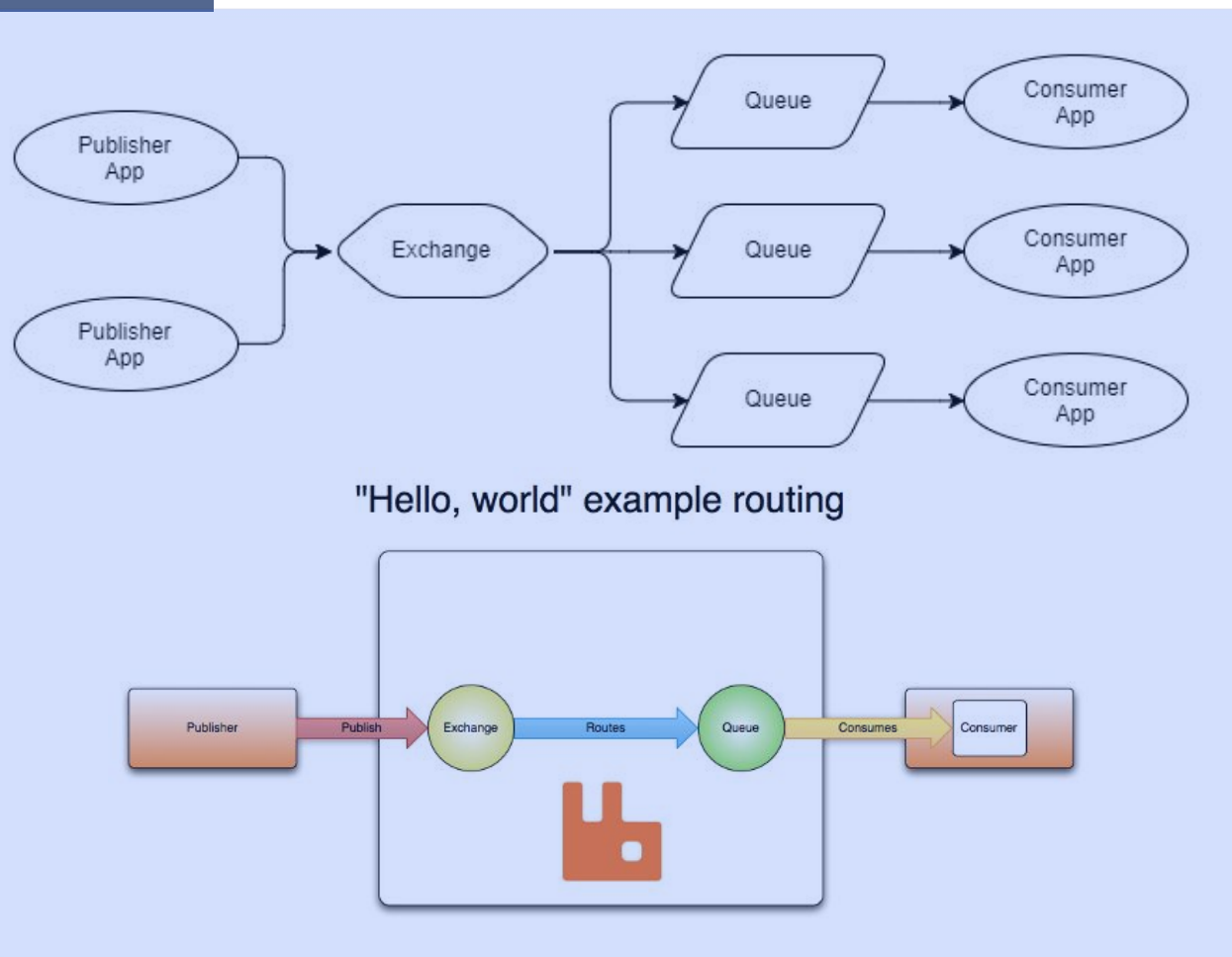
Группа получателей (Consumer group) – Получатели, подписанные на один и тот же топик.

Одна Партиция, один Блок, Один Получатель

Метод отдачи сообщений - Pull (получатель запрашивает данные)

ПОТОЧНЫЙ МЕТОД ПОЛУЧЕНИЯ ДАННЫХ

RabbitMQ



RabbitMQ — это распределенная система управления очередью сообщений.

Обзор:

Пабlishеры (publishers) отправляют сообщения на exchange'и

Exchange'и отправляют сообщения в очереди и в другие exchange'и

RabbitMQ отправляет подтверждения паблишерам при получении сообщения

Получатели (consumers) поддерживают постоянные TCP-соединения с RabbitMQ и объявляют, какую очередь(-и) они получают

RabbitMQ проталкивает (push) сообщения получателям

Получатели отправляют подтверждения успеха/ошибки

После успешного получения, сообщения удаляются из очередей

МЕТОД ЗАПРОСОВ К ДАННЫМ


SQL

SQL (*structured query language*) — декларативный информационно-логический язык программирования, применяемый для работы с данными хранящимися в реляционной базе данных.

SELECT — как показываем данные (определяем порядок и состав данных)

FROM — откуда берем данные
указывает на таблицу, по которой нужно делать запрос.

WHERE — какие данные показываем
Это фильтр **строк**, которые мы хотим вывести.
Например: вывести только те строки, где значение в колонке author — это “Dan Brown”.



```
1 SELECT *
2 FROM suppliers
3 WHERE (state = 'California' AND supplier_id <> 900)
4 OR (supplier_id = 100);
```

supplier_id	supplier_name	city	state
100	Microsoft	Redmond	Washington
200	Google	Mountain View	California
300	Oracle	Redwood City	California
700	Dole Food Compar	Westlake Village	California



bookid	title	author	published	stock
1	Scion of Ikshvaku	Amish Tripathi	06-22-2015	2
2	The Lost Symbol	Dan Brown	07-22-2010	3
3	Who Will Cry When You Die?	Robin Sharma	06-15-2006	4
4	Inferno	Dan Brown	05-05-2014	3
5	The Fault in our Stars	John Green	01-03-2015	3

МЕТОД ЗАПРОСОВ К ДАННЫМ

NoSQL



APIs

API (Application programming interface) — контракт, который предоставляет программа. «Ко мне можно обращаться так и эдак, я обязуюсь делать то и это».

REST (Representational State Transfer) - согласованный набор архитектурных принципов для создания масштабируемой и гибкой сети, описывающий компоненты системы и их взаимодействия)

RESTful —сеть, которая отвечает ограничениям Филдинга:

- Определение ресурсов
- Управление ресурсами через представления
- Гипермедиа
- Код по требованию
- Система слоёв
- Кэширование