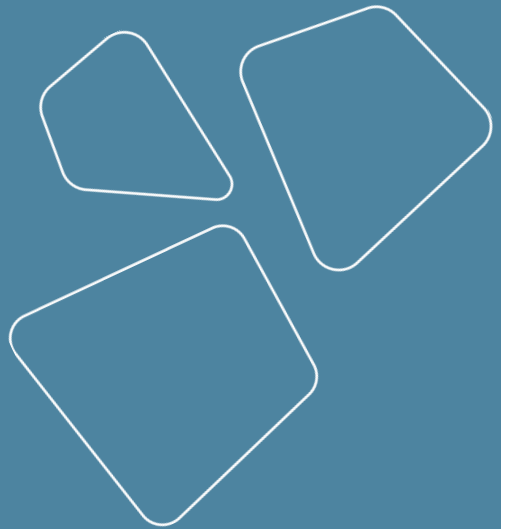
# Вычислительные ресурсы

Vladislav Goncharenko

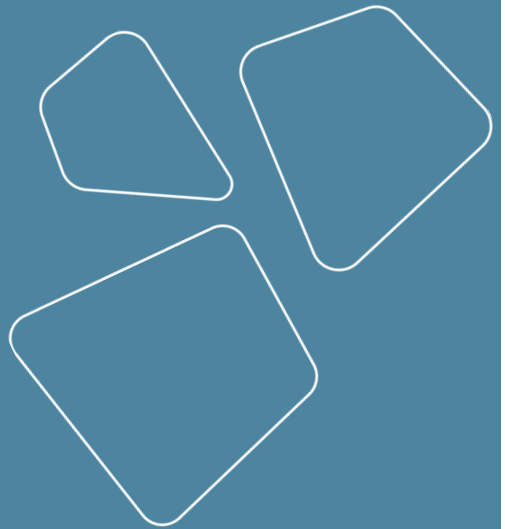
HSE, fall 2023

# Recap



- Логирование
- Визуализация
- Фреймворки для обучения

# Outline



- Контуры вычислений
  - Offline
  - Online (realtime)
  - Near real time
  - Edge
- Типы вычислительных ресурсов
- Jobs runners

# Контуры вычислений

girafe
ai

01

# Offline

Операции:

- могут занимать произвольное количество времени
- могут занимать произвольное количество ресурсов
- могут падать без последствий для бизнеса
- имеют относительно низкий приоритет выполнения

Примеры:

- Построение ежедневной аналитики
- EDA датасетов
- Обучение моделей

# Online (aka realtime)

*Real-time computing (RTC)* is the computer science term for hardware and software systems subject to a "real-time constraint", for example from event to system response. Real-time programs must guarantee response within specified time constraints, often referred to as "deadlines".

# Online (aka realtime)

Operations:

- must meet estimated time of arrival (ETA) (usually < 1 second)
- use predefined amount of resources
- failure causes apps to break
- have highest priority

Examples:

- site rendering
- quotes API of a stock market
- ML models inference
  - search engine results
  - recommender systems
  - advertisements
  - route generation on maps
  - chat bots

# Near real time

The term "near real-time" or "nearly real-time" (NRT), in telecommunications and computing, refers to the time delay introduced, by automated data processing or network transmission, between the occurrence of an event and the use of the processed data, such as for display or feedback and control purposes.

# Near real time

Operations:

- must meet estimated time of arrival (ETA) (usually < 10 minutes)
- use predefined amount of resources
- failure needs to be fixed with a bit relaxed deadline
- have high priority

Examples:

- site monitorings
- video processing
  - transcoding
  - video generation
- ALS models computing
- CI/CD processes

# Edge

Operations:

- must meet estimated time of arrival (ETA) (usually < 10-100 ms)
- use predefined amount of resources (stored on device)
- failure causes apps to break
- have high priority

Examples:

- IoT
- visual editors on smartphones
- bio-identification
- self-driving cars

# Типы вычислительных ресурсов

girafe ai

02

# Модель вычислений

Типы вычислительных мощностей

- Железные
- Виртуальные
  - classical VMs: KVM, vmware
  - docker
- Создаваемые под задачу (On-premises or vendor cloud)
  - MapReduce
  - Serverless computing (Amazon Lambda)
  - Очереди задач (slurm, clearml)
  - Kubernetes, k8s (Kubeflow)

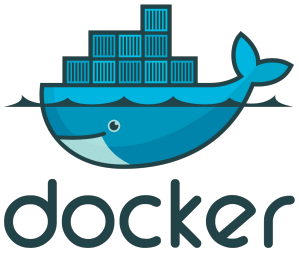
# Jobs runners

girafe
ai

03

# Job runners

- Docker Swarm
- Kubernetes (or k8s)
- очереди задач (slurm, clearml)
- Serverless (e.g. Амазон Лямбда)
- регулярные запуски джобов
- airflow
- cron




kubernetes

# Регулярные запуски кода

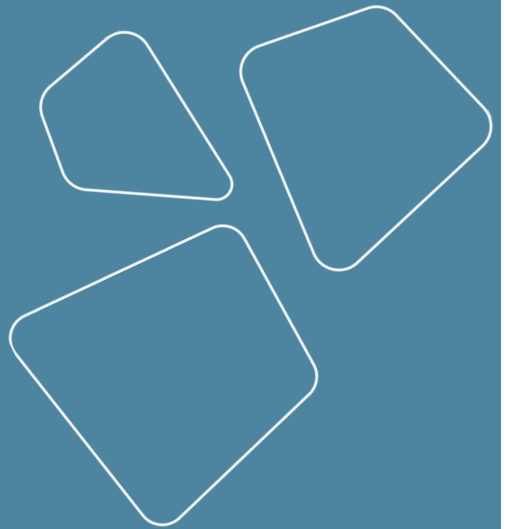
Дефакто стандарт индустри это airflow

Для других стеков применяются свои инструменты

Основная единица - Directed Acyclic Graph (DAG)

Apache Airflow

# О чём поговорили



- Контуры вычислений
  - Offline
  - Online (realtime)
  - Near real time
  - Edge
- Типы вычислительных ресурсов
- Jobs runners

# Спасибо за внимание!

Жду вопросов и обсуждений