Selectel

Инфраструктура для языковых моделей

Размер имеет значение

Владислав Кирпинский
Директор по облачной интеграции

Selectel сегодня

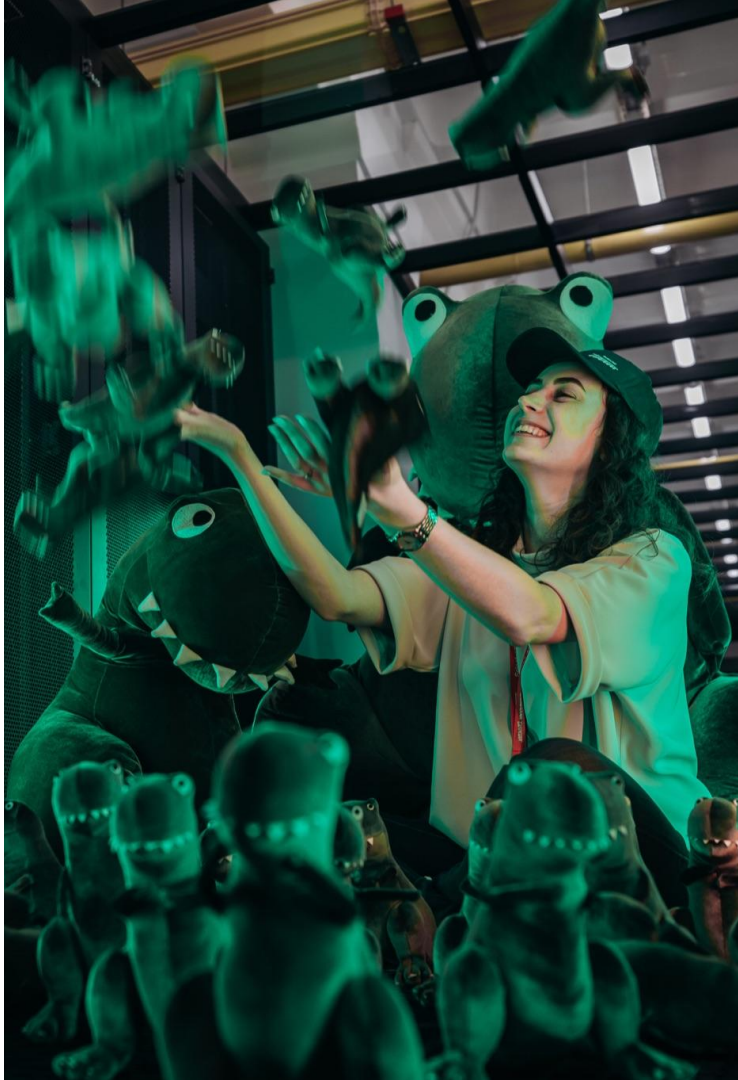
40+
продуктовых решений

6
собственных дата-центров

1 000+
сотрудников

24 000+
клиентов

С 2008 года помогаем компаниям решать бизнес-задачи, создавая надежную IT-инфраструктуру для проектов любой сложности.



Более 40 готовых инфраструктурных решений

Серверы и вычисления

- Выделенные серверы готовых и произвольных конфигураций
- Облачные серверы с моментальным запуском
- Серверы с GPU
- Экспериментальное железо

Облачная платформа

- Облачные базы данных
- Объектное хранилище
- Файловое хранилище
- Managed Kubernetes и Container Registry

Организация сети

- Сеть доставки контента (CDN)
- Балансировщики нагрузки
- Selectel Connect и сети L3 VPN
- Облачный DNS

Облако на базе VMware

- Публичное облако
- Частное облако
- Удаленные рабочие столы (VDI)

ML и обработка данных

- ML-платформа
- Платформа обработки данных
- Data Science & Analytical Virtual Machine

Безопасность

- Аттестованный сегмент ЦОД
- Соответствие 152-ФЗ
- Защита от DDoS и WAF

Единая панель управления и система биллинга

Документация к API и база знаний

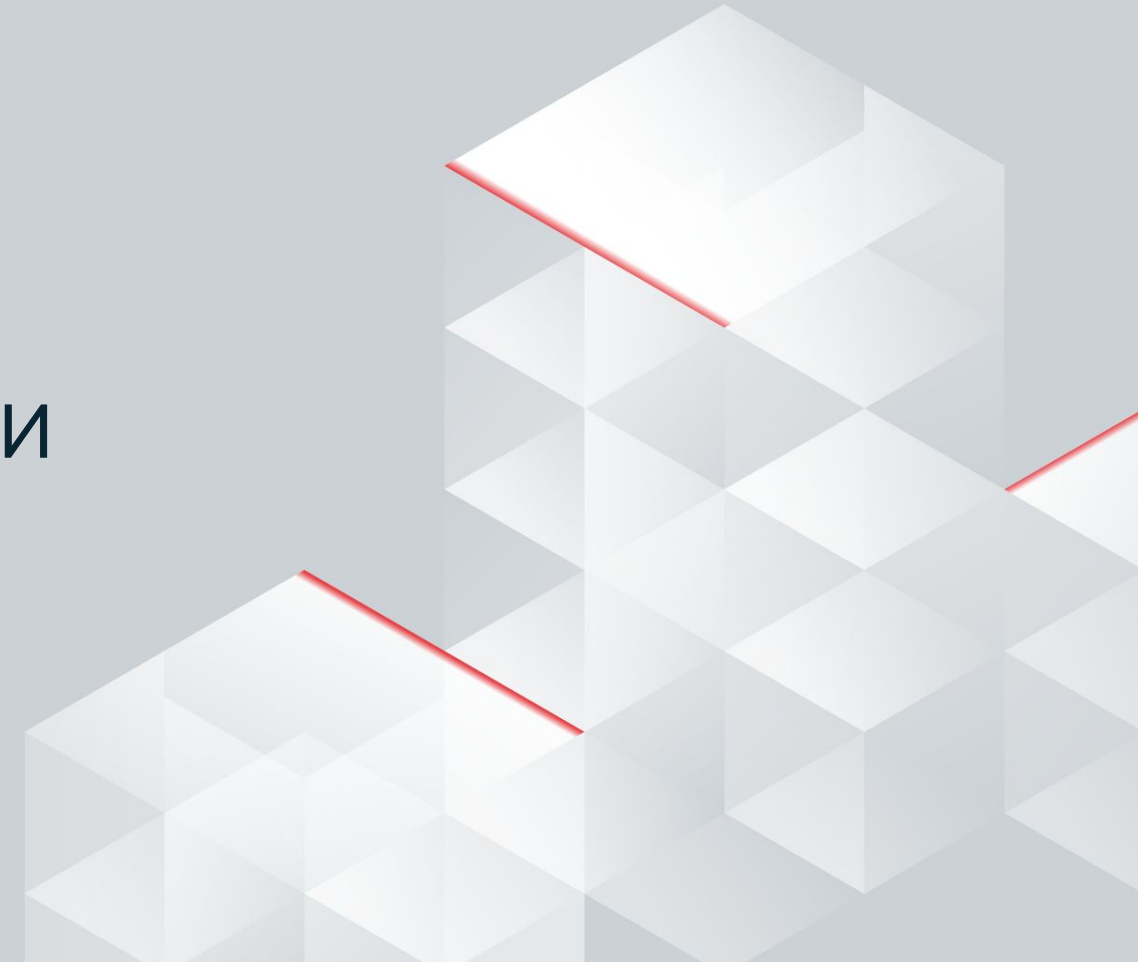
Система управления ролями (IAM)

Базовая защита от DDoS по умолчанию

Техническая поддержка 24/7

Содержание

Немного теории



Эй, Кембриджский словарь, что такое...

Inference

- A guess that you make or an opinion that you form based on the information that you have.

Training

- The process of learning the skills you need to do a particular job or activity.

Inference vs Training

IoT Data Input to ML Models

Raw IoT Data From IoT Endpoints

Logical Flow of Data

On-Premises or Cloud-Hosted

Training
Learning a New Capability From Existing Data

Deep-Learning Framework

Training Dataset

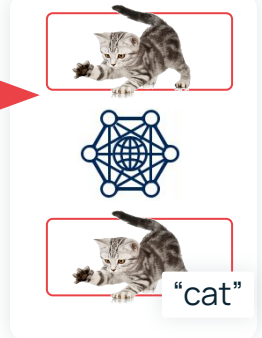
Logical Data Warehouse

Trained Model

Inference
Applying This Capability to New Data

Edge Device, On-Premises or Cloud-Hosted

New Data



App or Service Featuring Capability

Что такое NLP в 2024?

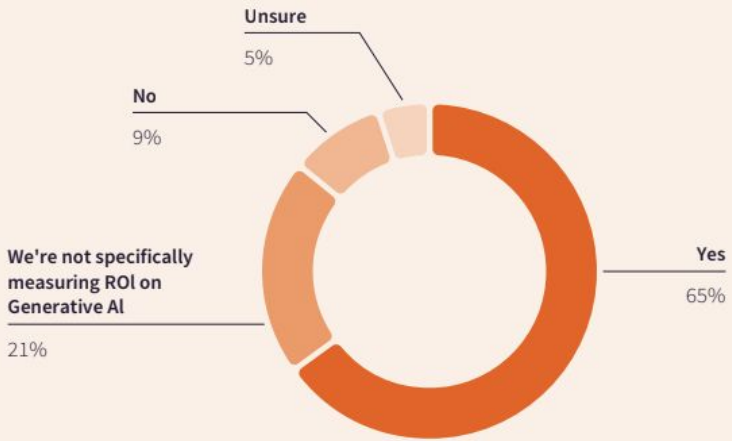
Large Language Models

ПОСЛЕ ДАННОГО СЛАЙДА ПРЕЗЕНТАЦИЯ В ПРОЦЕССЕ РЕДАКТИРОВАНИЯ

Автоматизация нагруженных процессов = экономия

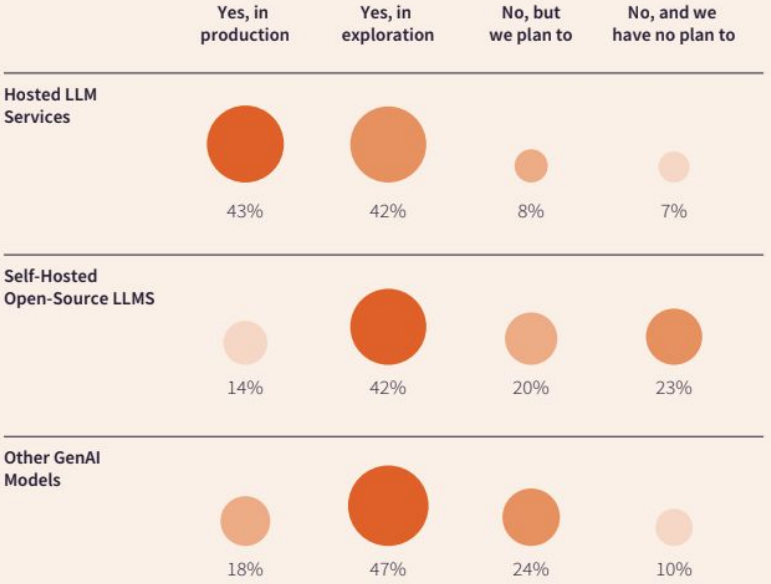
Experiencing Positive ROI From GenAI Use Cases

Q20new: Are you seeing positive ROI from GenAI use cases in production?



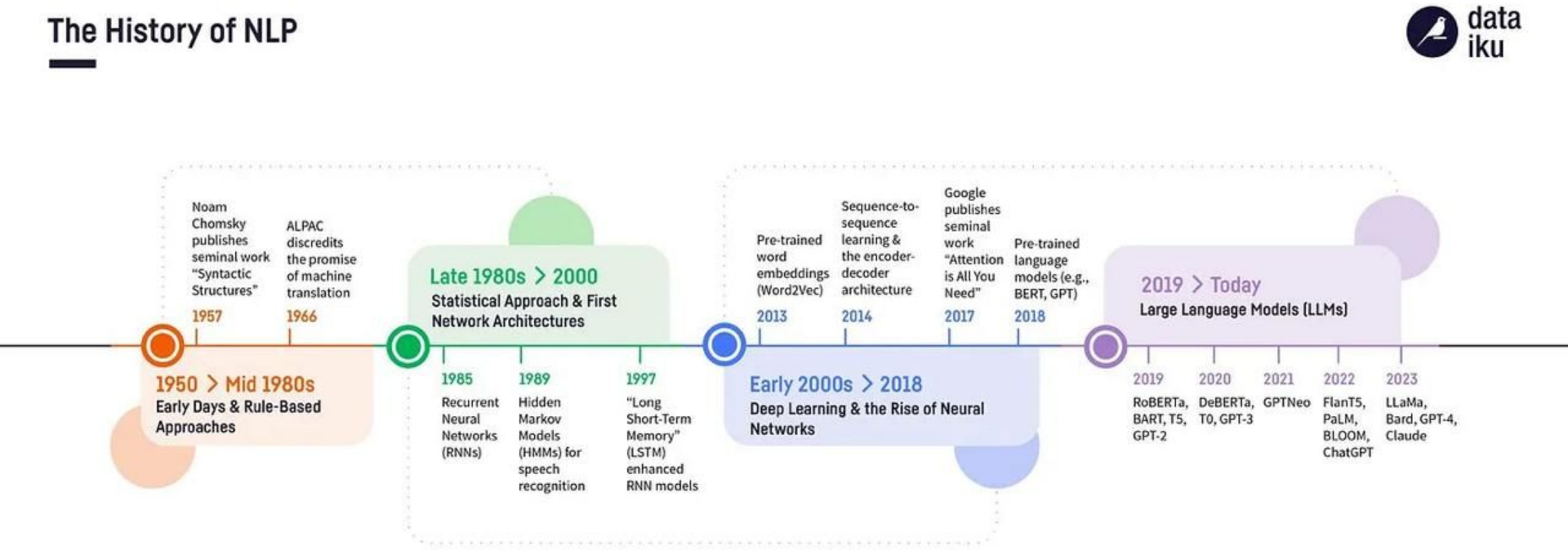
Type of GenAI Models / LLMs Currently Using

Q19new: Are you currently using any of the following GenAI models / LLMs?



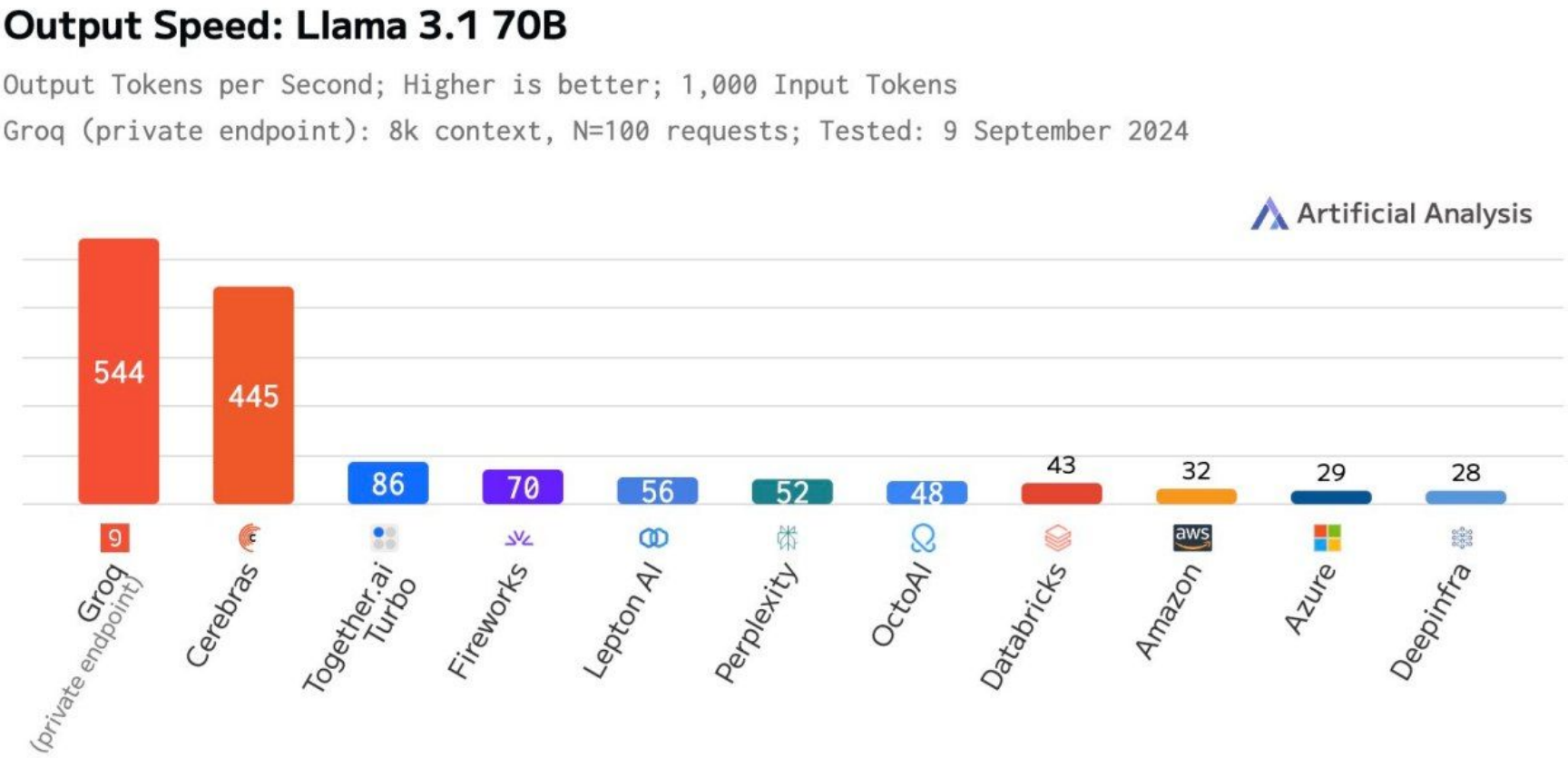
Источник

Но до них тоже все было неплохо...

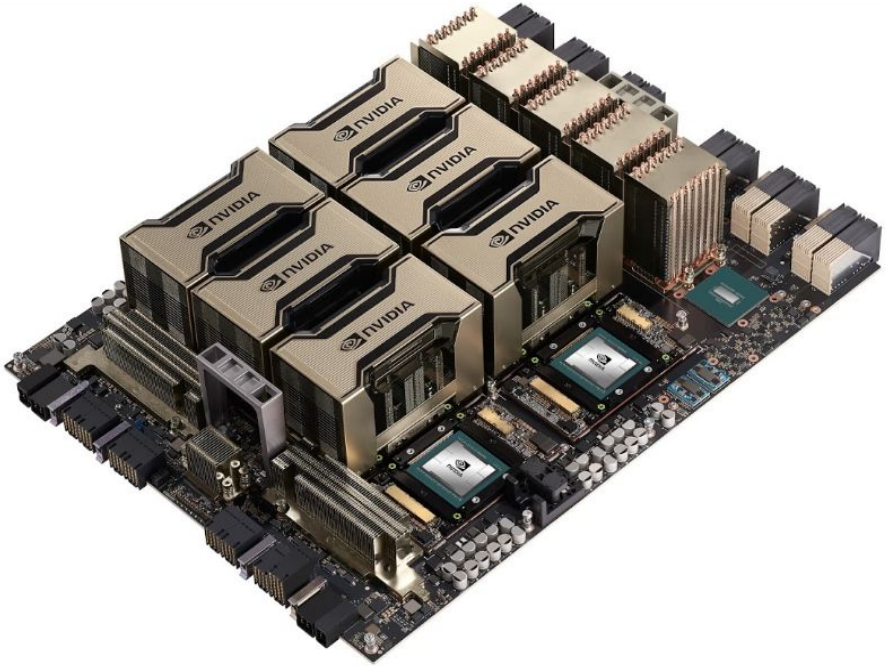


NLP-инфраструктура

Inference: битва за токены



ChatGPT: 3 500+ серверов NVIDIA HGX A100



Занимательный факт из реальной жизни

2 голосовых робота могут одновременно работать на **одном** ядре **CPU** с задержкой 1-2 секунды.

Источник

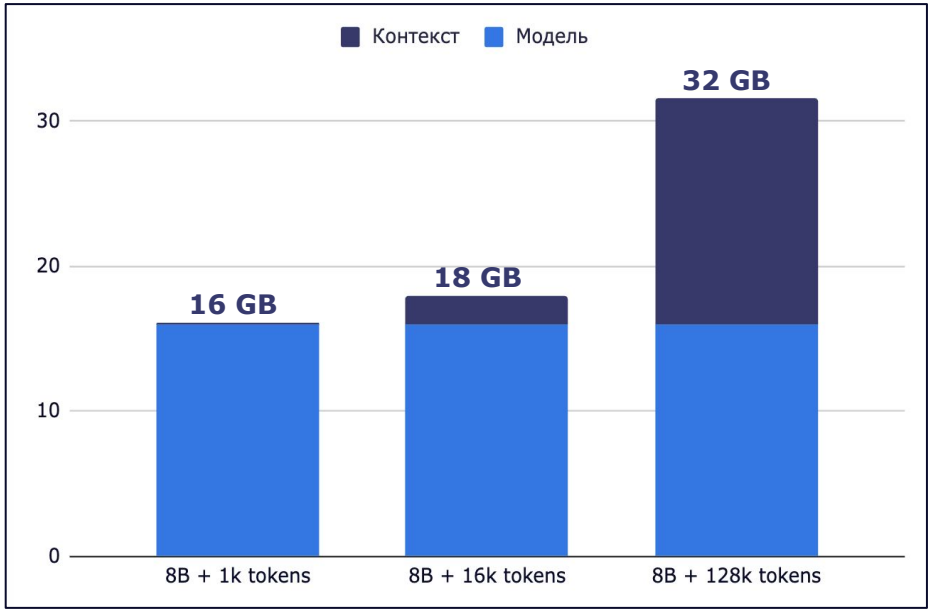
Training: битва за объем памяти

16,384
NVIDIA H100
80GB

Источник

Еще один занимательный факт из реальной жизни

Требования к памяти LLama 3.1 8B + контекст

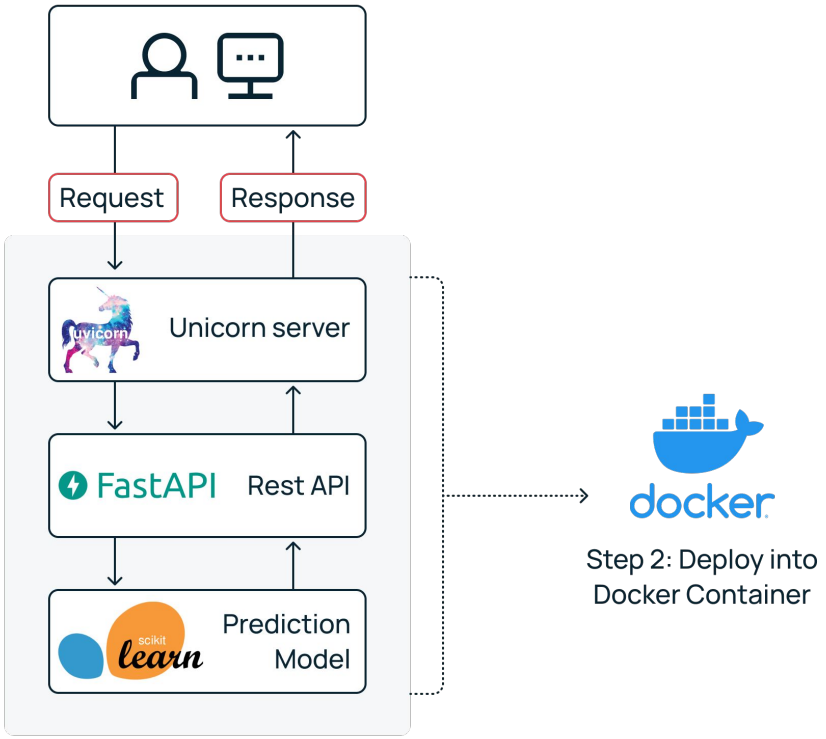


Для использования LLama 8B достаточно одной NVIDIA Tesla A100 40Gb

Источник

Inference-платформа

«Классический» путь ML-модели до Inference



Спекты задач

- Первый деплой
- Деплой новой версии
- Рост нагрузки
- Мониторинг
- Инференс-граф (LLM)

Внутри inference-платформы

S3

Managed K8S

Selectel
File Storage

Инфраструктура
Selectel

Istio

Prometheus
stack

Kiali

Дополнения

Системные компоненты

Ray

Triton

OpenSource решения

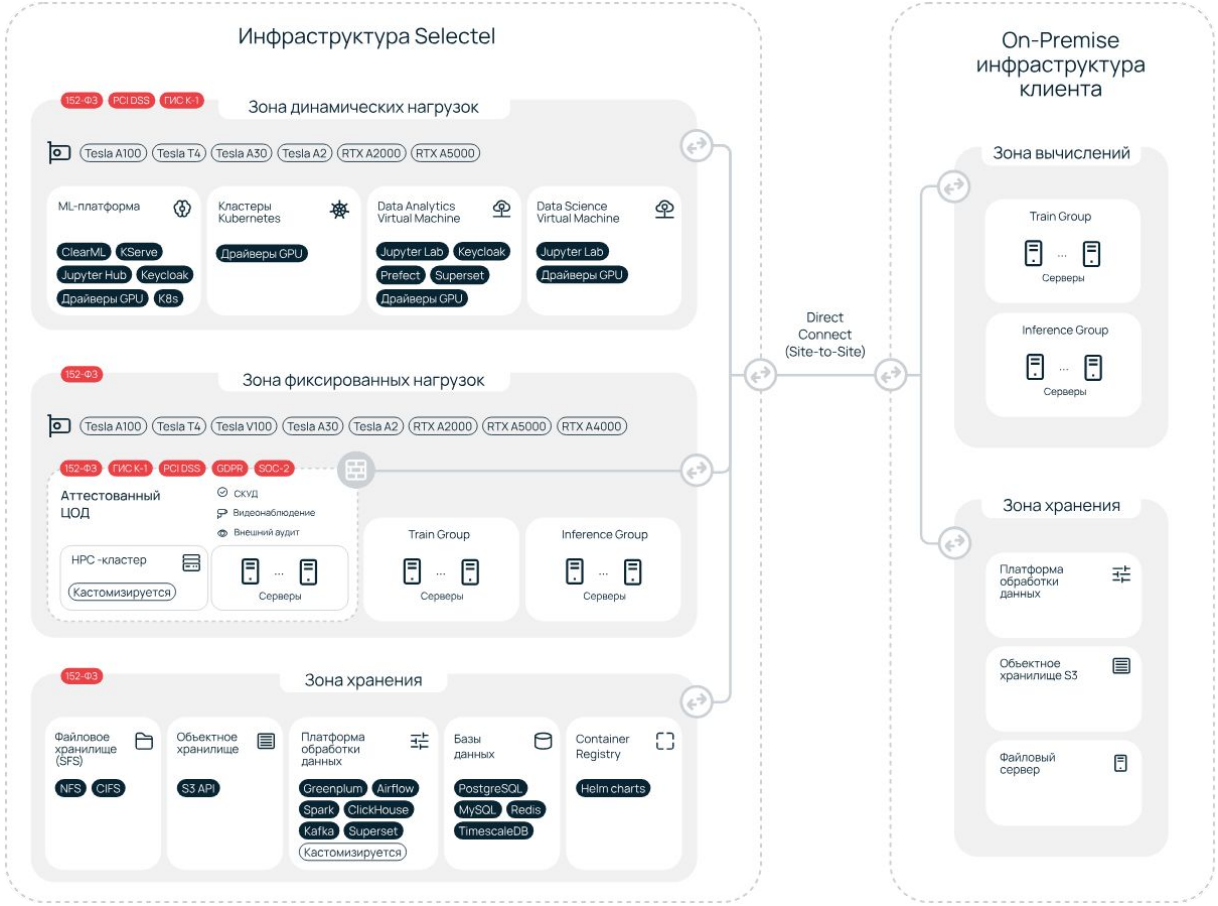
Инференс

Infrastructure чек-лист

Важно подумать над вопросами:

1. Какую NLP-технологию решено использовать?
2. Есть ли обоснование внедрения или нужен PoC?
3. Есть ли ответственные от IT и целевого бизнес-подразделения?
4. Какова предполагаемая нагрузка на итоговый сервис?
5. Нужно ли дообучать ML-модели?
6. Подходит ли формат собираемых бизнес-данных для дообучения ML-моделей?
7. Нужны ли изолированные контуры для training и inference?
8. Где провести тесты разных GPU под задачу?
9. Купить GPU-серверы или арендовать?
10. Требуется ли соблюдать 152-ФЗ?
11. Какой стек инструментов будет использоваться?
12. Есть ли необходимые компетенции?

Selectel — провайдер для ML



Selectel

selectel.ru

selectel

selectel

Остались вопросы? Обращайтесь!