

ПРИЛОЖЕНИЕ №1.EVO.11.1
к Договору**ОПИСАНИЕ И УСЛОВИЯ ПРЕДОСТАВЛЕНИЯ УСЛУГИ**
«EVOLUTION ML INFERENCE»**1. ОБЩАЯ ИНФОРМАЦИЯ И ОПИСАНИЕ УСЛУГИ**

- 1.1. Evolution ML Inference (далее – Услуга) – это облачное решение для развертывания ML-моделей, которое поддерживает динамическое масштабирование и взаимодействие с разными источниками событий, такими как HTTP-запросы.
- 1.2. Услуга реализована на оборудовании, принадлежащем Исполнителю, и средствами системы виртуализации собственной разработки (в т. ч. на базе компонентов с открытым исходным кодом). Ресурсом Услуги является ML-модель.
- 1.3. Функциональные возможности:
- 1.3.1. развертывание ML-моделей для обработки запросов или событий;
 - 1.3.2. развертывание контейнеров для обработки запросов или событий;
 - 1.3.3. загрузка моделей из HuggingFace публичных и частных репозиторий;
 - 1.3.4. автоматическое масштабирование модели в зависимости от нагрузки;
 - 1.3.5. управление и контроль доступа к модели;
 - 1.3.6. мониторинг и сбор статистики по потреблению ресурсов модели;
 - 1.3.7. управление конфигурацией модели;
- 1.4. Услуга состоит из следующих компонентов:
- 1.4.1. сервис-контроллер для управления Услугой из личного кабинета Облачной Платформы;
 - 1.4.2. компонент сбора данных мониторинга;
 - 1.4.3. компонент, отвечающий за управление жизненным циклом продукта;
 - 1.4.4. компонент, отвечающий за сбор тарификационных данных;
 - 1.4.5. интеграция с сервисом Evolution Artifact Registry;
 - 1.4.6. Платформа Evolution, обеспечивающая выбор виртуальных машин для разворачивания программного обеспечения;
- 1.5. Обеспечение защиты инфраструктуры.
- Защита инфраструктуры Облачной Платформы обеспечивается на следующих уровнях:
- на физическом уровне;
 - на сетевом уровне;
 - на инфраструктурном уровне;
 - обеспечение защиты от несанкционированного доступа к виртуальным машинам, на которых функционирует сервис;
 - антивирусная защита виртуальных машин, на которых функционирует сервис;
 - периодическая проверка на соответствие требованиям информационной безопасности (в том числе с использованием сканеров безопасности) и обновление образов виртуальных машин, используемых сервисом, и установленной на них операционной системой;
 - мониторинг и реагирование на инциденты информационной безопасности, возникающие при функционировании сервиса;
 - межсетевое экранирование сетевых потоков сервиса средствами Платформы Evolution.

2. БАЗОВАЯ ФУНКЦИОНАЛЬНОСТЬ И ПАРАМЕТРЫ РЕСУРСОВ УСЛУГИ

- 2.1. Параметры Ресурсов Услуги:

Таблица 2. Параметры предоставляемых Ресурсов

Услуга	Тарифицируемые Ресурсы	Характеристики и метрики	Допустимые значения
Shared GPU ¹	Видео память GPU H100 PCIe	Объем vRAM GPU (ГБ)	1
		Время работы (минуты)	1
	Видео память GPU A100 PCIe	Объем vRAM GPU (ГБ)	1
		Время работы (минуты)	1
	Видео память GPU V100 NVLink	Объем vRAM GPU (ГБ)	1
		Время работы (минуты)	1

¹ Shared GPU – Заказчику предоставляется возможность частичного потребления GPU-ресурса (в пределах его физического объема памяти), что позволяет гибко утилизировать ресурсы, а также эффективно (с т.з. цены) размещать ML-модели.

Вычислительные ресурсы	Видео память GPU H100 NVLink	Объем vRAM GPU (ГБ)	1
		Время работы (минуты)	1
	Инстанс тип 1xH100 NVLink /20vCPU/190Gb RAM	Количество vCPU (шт.)	20
		Объём RAM (ГБ)	190
		Количество GPU (шт.)	1
		Время работы (минуты)	1
	Инстанс тип 2xH100 NVLink /40vCPU/380Gb RAM	Количество vCPU (шт.)	40
		Объём RAM (ГБ)	380
		Количество GPU (шт.)	2
		Время работы (минуты)	1
	Инстанс тип 4xH100 NVLink /80vCPU/760Gb RAM	Количество vCPU (шт.)	80
		Объём RAM (ГБ)	760
		Количество GPU (шт.)	4
		Время работы (минуты)	1
	Инстанс тип 6xH100 NVLink /120vCPU/1140Gb RAM	Количество vCPU (шт.)	120
		Объём RAM (ГБ)	1 140
		Количество GPU (шт.)	6
		Время работы (минуты)	1
	Инстанс тип 8xH100 NVLink /160vCPU/1520Gb RAM	Количество vCPU (шт.)	160
		Объём RAM (ГБ)	1 520
		Количество GPU (шт.)	8
		Время работы (минуты)	1
	Инстанс тип 1xV100 NVLink /4vCPU/64Gb RAM	Количество vCPU (шт.)	4
		Объём RAM (ГБ)	64
		Количество GPU (шт.)	1
		Время работы (минуты)	1
	Инстанс тип 2xV100 NVLink /8vCPU/128Gb RAM	Количество vCPU (шт.)	8
		Объём RAM (ГБ)	128
		Количество GPU (шт.)	2
		Время работы (минуты)	1
	Инстанс тип 4xV100 NVLink /16vCPU/256Gb RAM	Количество vCPU (шт.)	16
		Объём RAM (ГБ)	256
		Количество GPU (шт.)	4
		Время работы (минуты)	1
	Инстанс тип 8xV100 NVLink /32vCPU/512Gb RAM	Количество vCPU (шт.)	32
		Объём RAM (ГБ)	512
		Количество GPU (шт.)	8
		Время работы (минуты)	1
	Инстанс тип 16xV100 NVLink/64vCPU/1024Gb RAM	Количество vCPU (шт.)	64
		Объём RAM (ГБ)	1024
		Количество GPU (шт.)	16
		Время работы (минуты)	1
	Инстанс тип 1xA100 PCIe/20vCPU/125Gb RAM	Количество vCPU (шт.)	20
		Объём RAM (ГБ)	125
		Количество GPU (шт.)	1
		Время работы (минуты)	1
	Инстанс тип 2xA100 PCIe/40vCPU/250Gb RAM	Количество vCPU (шт.)	40
		Объём RAM (ГБ)	250
		Количество GPU (шт.)	2
		Время работы (минуты)	1
	Инстанс тип 4xA100 PCIe/80vCPU/500Gb RAM	Количество vCPU (шт.)	80
		Объём RAM (ГБ)	500
		Количество GPU (шт.)	4
		Время работы (минуты)	1
	Инстанс тип 6xA100 PCIe/120vCPU/750Gb RAM	Количество vCPU (шт.)	120
		Объём RAM (ГБ)	750
		Количество GPU (шт.)	6
		Время работы (минуты)	1
	Инстанс тип 8xA100 PCIe/160vCPU/1000Gb RAM	Количество vCPU (шт.)	160
		Объём RAM (ГБ)	1 000
		Количество GPU (шт.)	8
		Время работы (минуты)	1
	Инстанс тип 1xH100 PCIe/20vCPU/125Gb RAM	Количество vCPU (шт.)	20
		Объём RAM (ГБ)	125
		Количество GPU (шт.)	1
		Время работы (минуты)	1
	Инстанс тип 2xH100 PCIe/40vCPU/250Gb RAM	Количество vCPU (шт.)	40
		Объём RAM (ГБ)	250
		Количество GPU (шт.)	2
		Время работы (минуты)	1

	Инстанс тип 4xH100 PCIe/80vCPU/500Gb RAM	Количество vCPU (шт.)	80
		Объём RAM (ГБ)	500
		Количество GPU (шт.)	4
		Время работы (минуты)	1
	Инстанс тип 6xH100 PCIe/120vCPU/750Gb RAM	Количество vCPU (шт.)	120
		Объём RAM (ГБ)	750
		Количество GPU (шт.)	6
		Время работы (минуты)	1
	Инстанс тип 8xH100 PCIe/160vCPU/1000Gb RAM	Количество vCPU (шт.)	160
		Объём RAM (ГБ)	1 000
		Количество GPU (шт.)	8
		Время работы (минуты)	1
Кэш ML-моделей ²	Хранение модели	Объём модели (ГБ)	1
		Время работы (минуты)	1
Запросы к ML-моделям	Запросы в запущенный Инстанс	Запросы (шт)	1 000 000

3. ТАРИФИКАЦИЯ УСЛУГИ

- 3.1. Для данной Услуги используется Динамическая тарификация (Pay as you go). Клиент начинает платить за запущенную модель после переход ее в статус «Запущено», и плата начисляется за потребляемые вычислительные ресурсы, хранения модели и количеству обращений в модель.
- 3.2. Динамическая тарификация предполагает оплату пула Ресурсов (см. п. 2.1. Приложения) по факту их потребления Заказчиков в течение Отчетного периода.
- 3.3. Окончательная стоимость Услуги в Отчётном периоде формируется в соответствии с тарифами, установленными в Приложении №7.EVO.11.1 к Договору.

3.4. Объекты тарификации:

- Тарифицируются вычислительные ресурсы;
- Тарифицируется хранение модели в кэше;
- Тарифицируются запросы к модели.

- 3.5. Величина ежемесячного платежа за пользование Услугой определяется в соответствии с фактическим потреблением Ресурсов. Доступные Ресурсы и методика расчета перечислены в примере ниже:

3.6. Пример расчета

3.6.1. Для Shared GPU¹

- Общая формула расчета:

$$\text{Стоимость} = (vRAM \text{ Гб} * \text{цена 1Гб } vRAM \text{ GPU} + (\text{Запросы (в млн.)} * 12.8 \text{ Р/миллион запросов}) + \text{Кэш ML} - \text{модели(Гб)} * 0.013 \text{ Р/Гб}) * \left(\frac{\text{Время в секундах}}{3600} \right)$$

Где: vRAM Гб – объем выделенной видеопамяти GPU в гигабайтах;

цена 1Гб vRAM GPU – стоимость 1Гб видео памяти GPU карты, указана в Тарифах Услуги;

Запросы – количество обработанных запросов (в миллионах);

Кэш ML-модели(Гб) – объем модели в кэше в гигабайтах;

Время в секундах – продолжительность работы в секундах.

- Пример расчета (Цена 1Гб H100: 5,625 руб; Запросы: 5 млн; Объем модели: 4 Гб; Время: 1 час.):

$$\text{Стоимость} = (8 * 5.625 + (5 * 12.8) + 4 * 0.013) * (3600/3600) = 109.052$$

3.6.2. Для Инстанс типов:

- Общая формула расчета:

$$\text{Стоимость} = \text{стоимость Инстанса в секунду} + \text{Запросы (в млн.)} * 12.8 \frac{\text{Р}}{\text{миллион}} \text{запросов} + \text{Кэш ML} - \text{модели(Гб)} * 0.013 \frac{\text{Р}}{\text{Гб}} * \frac{\text{Время в секундах}}{3600}$$

Где:

Стоимость Инстанса - фиксированная стоимость выделенного оборудования, указана в Тарифах Услуги;

Запросы – количество обработанных запросов (в миллионах);

² Кэш ML-моделей: временные файлы, формируемые запущенной ML-Моделью, необходимые для ее работы. Указанное пространство не является хранилищем Заказчика (в т.ч. для долгосрочного хранения информации), очищается автоматически в момент, когда ML-Модель не используется Заказчиком.

Кеш ML-модели(Гб) – объем модели в кэше в гигабайтах;

Время в секундах – продолжительность работы в секундах.

- Пример расчета (Стоимость Инстанса: 0.125 ₽/сек; Запросы: 5 млн; Объем модели: 20 Гб; Время: 1 час.):

$$\text{Стоимость} = 0.125 * 3600 + 5 * 12.8 + 20 * 0.013 * (3600/3600) = 514.26 \text{ Р}$$

4. ИНЫЕ УСЛОВИЯ, ПРИМЕНИМЫЕ К УСЛУГЕ

- 4.1. Возможные виды подключения / изменения / отключения Услуги:
 - 4.1.1. Посредством совершения действий в Личном Кабинете.
 - 4.1.2. В отношении с GPU – в порядке, установленном в пункте п.4.5 Приложения.
- 4.2. Возможный порядок расчётов по Услуге:
 - Предоплата³;
 - Постоплата⁴.
- 4.3. Возможные способы оплаты / порядок пополнения баланса:
 - 4.3.1. В безналичном порядке на основании выставленного Исполнителем счёта;
 - 4.3.2. оплата посредством электронных средств платежа.
- 4.4. Требования к инфраструктуре Заказчика:
 - 4.4.1. Наличие доступа в Интернет.
- 4.5. Стороны установили следующий порядок Заказа GPU/Увеличения объема памяти GPU по Приложению:
 - 4.5.1. Подключение Услуги осуществляется Исполнителем на основании Запроса на изменение (ЗНИ) через службу технической поддержки Исполнителя. Запрос должен быть направлен не позднее, чем за 6 (шесть) рабочих дней до желаемой даты начала потребления Услуги;
 - 4.5.2. В течение 3 (трех) рабочих дней Исполнитель обязуется рассмотреть ЗНИ на подключение Услуги и направить ответ (информацию о подключении Услуги или отказ в её предоставлении Услуги);
 - 4.5.3. В случае согласования Сторонами Заказа Услуги она предоставляется в дату начала её оказания (в соответствии с информацией в ЗНИ) с 10:00 по московскому времени.

5. ОСОБЕННОСТИ УРОВНЯ ПРЕДОСТАВЛЕНИЯ УСЛУГИ

- 5.1. В соответствии с пп. 1.1.4. вносятся следующие уточнения в уровень предоставления Услуги, действующий в отношении услуг Evolution по умолчанию (Приложения №2.EVO.0. к Договору).
- 5.2. Для Услуги устанавливаются следующие особенности определения уровня Доступности:
 - 5.2.1. Доступность рассчитывается отдельно для каждого Ресурса Услуги (п. 1.2. Приложения);
 - 5.2.2. Недоступностью Услуги является ситуация, при которой ML-моделей была развернута и, находясь в статусе "Запущена", не принимает запросы/события и не дает ответа в течение 5 (пяти) и более минут по причинам, зависящим от Cloud.ru
 - 5.2.3. Компенсация выплачивается пропорционально объёму недоступных Ресурсов Услуги, т.е. Компенсация за нарушение целевых показателей Доступности Услуги рассчитывается согласно количеству недоступных Ресурсов.
- 5.3. Во всём остальном в части уровня предоставления Услуги применимы положения Приложения №2.EVO.0. к Договору.

³ Является способом по умолчанию для физических лиц, присоединившихся к условиям Договора/Оферте путём акцепта (п. 1.5. Договора/Оферты).

⁴ Является способом по умолчанию для юридических лиц.