ПРИЛОЖЕНИЕ №1.EVO.11.
к Договору

ОПИСАНИЕ И УСЛОВИЯ ПРЕДОСТАВЛЕНИЯ УСЛУГИ «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. Функциональные возможности:
1.2.1. развертование ML-моделей для обработки запросов или событий;
1.2.2. развертование контейнеров для обработки запросов или событий;
1.2.3. загрузка моделей из HuggingFace публичных и приватных репозиториев;
1.2.4. автоматическое масштабирование модели в зависимости от нагрузки;
1.2.5. управление и контроль доступа к модели;
1.2.6. мониторинг и сбор статистики по потреблению ресурсов модели;
1.2.7. управление конфигурацией модели;

1.3. Услуга состоит из следующих компонентов:
1.3.1. сервис-контроллер для управления Услугой из личного кабинета Облачной Платформы;
1.3.2. компонент сбора данных мониторинга;
1.3.3. компонент, отвечающий за управление жизненным циклом продукта;
1.3.4. компонент, отвечающий за сбор тарификационных данных;
1.3.5. интеграция с сервисом Evolution Artifact Registry;
1.3.6. Платформа Evolution, обеспечивающая выбор виртуальных машин для разворачивания программного обеспечения;

1.4. 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.

1.5. Квоты и ограничения, которые накладываются на запуск моделей Заказчика в рамках одной Организации, описаны в Таблице 1.

Таблица 1. Ограничения по доступным объёмам услуг в рамках Организации

Объекты	Единицы	Ограничения
Видео память GPU H100 PCIe	Гб	0
Видео память GPU A100 PCIe	Гб	0
Видео память GPU V100 NVLink	Гб	12
Видео память GPU H100 NVLink	Гб	0
GPU	Шт	0

2. БАЗОВАЯ ФУНКЦИОНАЛЬНОСТЬ И МЕТРИКИ УСЛУГИ

2.1. Параметры Услуги:

Таблица 2. Параметры предоставляемых услуг

Услуга	Тарифицируемые единицы	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
	Видео память GPU H100 NVLink	Объем vRAM GPU (ГБ)	1
		Время работы (минуты)	1
Вычислительные ресурсы	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

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

		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	Объем модели (Гб)	1
		Время работы (минуты)	1
Запросы к ML-моделям	Запросы в замущеный 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 ресуры, 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. к Договору.

3.4. Объекты тарификации:

- Тарифицируются вычислительне 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 расчет

3.6.1. Для Shared GPU[1]

- Общая формула расчета:

$$\text{Стоимость} = (vRAM\ \text{Гб} * \text{цена 1Гб}\ vRAM\ GPU + (\text{Запросы (в млн.)} * 12.8\ ₽/\text{миллион запросов}) + \text{Кеш}\ ML - \text{модели(Гб)} * 0.013\ ₽/\text{Гб}) * \text{Время в часах}$$

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).;
Кеш ML-модели(Гб) – Объем модели в кэше в гигабайтах;
Время в часах – Продолжительность работы в часах.

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

$$\text{Стоимость} = (8 * 5.625 + (5 * 12.8) + 4 * 0.013) * 1 = 109.052$$

[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.

3.6.2. Для Инстанс типов:

- Общая формула расчета:

$$\text{Стоимость} = \text{Стоимость Инстанса в час} + \text{Запросы (в млн.)} * 12.8\ ₽/\text{миллион запросов} + \text{Кеш } ML\text{-модели(Гб)} * 0.013\ ₽/\text{гб} * \text{Время в часах}$$

Где:

Стоимость Инстанса - Фиксированная стоимость выделенного оборудования указана в Тарифах Услуги;

Запросы– Количество обработанных запросов (в миллионах).;

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

Время в часах – Продолжительность работы в часах.

- Пример расчета (Стоимость инстанса: 450 ₽/час; Запросы: 5 млн; Объем модели: 20 ГБ; Время: 1 час.):

$$\text{Стоимость} = 450 + 5 * 12.8 + 20 * 0.013 * 1 = 514.26\ ₽$$

4. ДОСТУПНОСТЬ УСЛУГ

4.1. Показатели доступности Evolution ML Inference:

Таблица 2. Показатели доступности Услуги

Наименование услуги	Доступность Услуги за Отчетный период, %
Evolution ML Inference	99,9%

5. ИНЫЕ УСЛОВИЯ, ПРИМЕНИМЫЕ К УСЛУГЕ

5.1. Возможные виды подключения / изменения / отключения Услуги:

5.1.1. Посредством совершения действий в Личном Кабинете.

5.1.2. В отношении с GPU — в порядке, установленном в пункте п.5.5 Приложения

5.2. Возможный порядок расчётов по Услуге:

- Предоплата[3];
- Постоплата[4];

5.3. Возможные способы оплаты / порядок пополнения баланса:

5.3.1. В безналичном порядке на основании выставленного Исполнителем счёта;

5.3.2. оплата посредством электронных средств платежа.

5.4. Требования к инфраструктуре Заказчика:

5.4.1. Наличие доступа в Интернет.

5.5. Стороны установили следующий порядок Заказа GPU/Увеличиения объема памяти GPU по Приложению:

5.5.1. 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;

5.5.2. 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);

5.5.3. 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.

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

[4] Является способом по умолчанию для юридических лиц