ARENADATA
Корпоративная платформа хранения
и обработки больших данных

Общество с ограниченной ответственностью
«Аренадата Софтвер»

Описание технической архитектуры программного обеспечения для электронно-вычислительных машин Arenadata Enterprise SQL Umformer (ADESU)

Москва
2025

Термины и определения

Термин	Значение
ANSI SQL	Стандартный язык запросов к базам данных, разработанный Американским национальным стандартом (ANSI)
LLM-модели	Языковые модели, обученные на больших объемах текстовых данных

Сокращения и обозначения

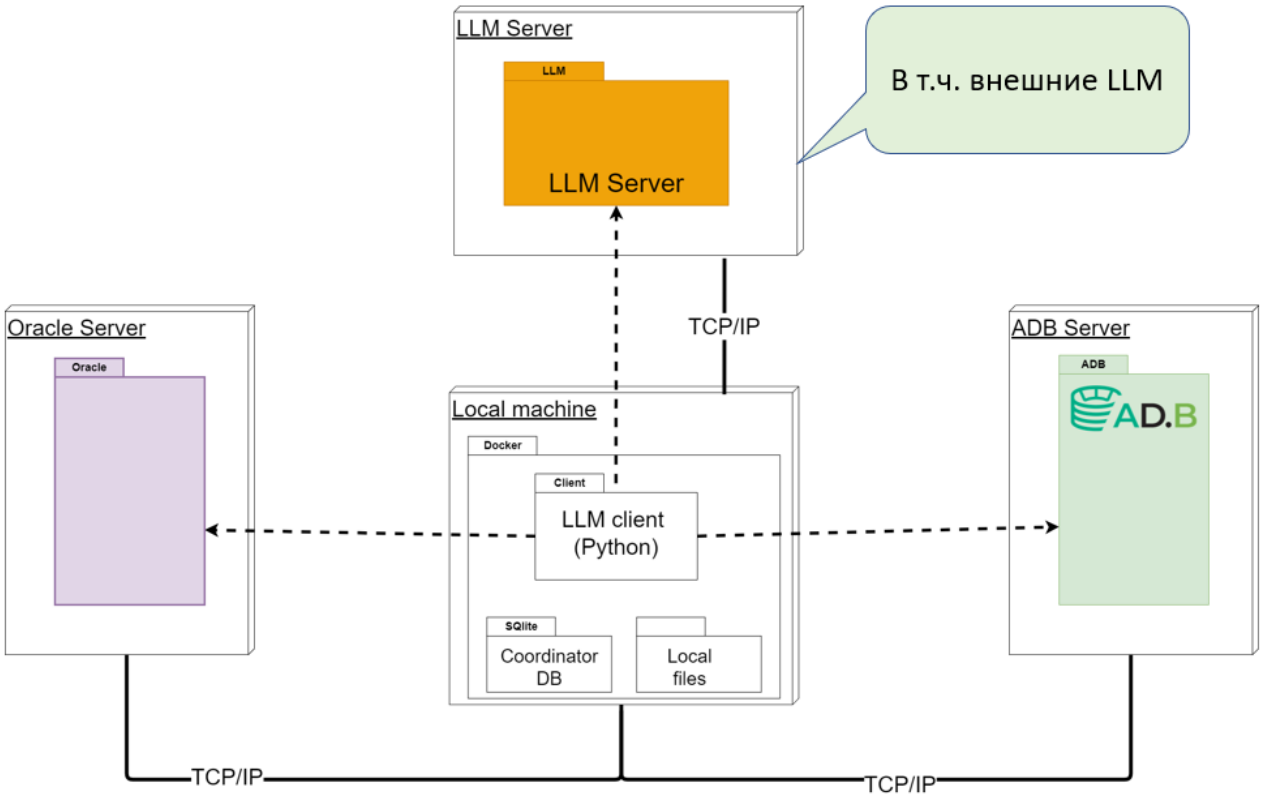
Сокращение	Наименование
ADB	Arenadata DB
ADESU, ПО Umformer	Arenadata Enterprise SQL Umformer
API	(англ. – Application Programming Interface) – набор классов, процедур, функций, структур или констант, которыми одна компьютерная программа может взаимодействовать с другой программой
CPU	(англ. – Central Processing Unit) – центральный процессор
IP	(англ. – Internet Protocol) – маршрутизируемый протокол сетевого уровня стека TCP/IP
LLMs	Large language models
MSSQL	Microsoft SQL Server
SQL	(англ. – Structured Query Language) – декларативный язык программирования, применяемый для создания, модификации и управления данными в реляционных базах данных
БД	База данных
ПО	Программное обеспечение
РФ	Российская Федерация
СУБД	Система управления базами данных

Arenadata Enterprise SQL Umformer (ADESU) — это ПО, предназначенное для автоматического преобразования кода из одной системы в другую. ПО использует LLM-модели для трансформации запросов и обеспечивает проверку синтаксиса полученных запросов. Основные функции ПО:

1. Трансформация sql-запросов с помощью LLM (источником и приемником могут быть следующие СУБД: Oracle, ADB, MSSQL, Impala, Hive);
2. Трансформация функций и процедур Oracle в функции ADB с помощью LLM;
3. Трансформация функций и процедур Oracle в функции ADB с помощью ora2pg.

Решение использует локальные open-source модели LLM и связанные технологии, такие как:

1. Ollama
2. vLLM
3. Langchain



Верхнеуровневая архитектурная схема ADESU

Ollama – это фреймворк, который позволяет запускать языковые модели локально на различных устройствах. Он объединяет модельные веса и окружение в приложение, которое работает на устройстве и обслуживает языковую модель. Ollama поддерживает количественное снижение точности весов модели, чтобы уменьшить объем памяти, необходимый для хранения языковой модели в памяти. Это особенно важно при использовании потребительского оборудования, такого как CPU или ноутбук с графическим процессором. Ollama также предлагает различные версии моделей, которые можно скачать через команду `ollama pull`.

Фреймворк Ollama создан для того, чтобы облегчить использование больших языковых моделей (LLM) на локальных устройствах. Ключевые технические характеристики и особенности Ollama:

1. Поддерживаемые устройства:
    - Локальные компьютеры (CPU/GPUs);
    - Серверы.
2. Языковые модели:
    - Поддерживаются различные LLMs, включая GPT-J, MPT и другие.
3. Форматы данных:
    - Веса моделей могут быть загружены в формате *.bin* или *.pt*.
4. Интеграция с API:
    - Предоставляет RESTful API для взаимодействия с моделью.
5. Оптимизация производительности:
    - Возможность использования квантования для уменьшения объема памяти, необходимой для хранения весов модели;
    - Адаптивная подстройка параметров модели для повышения скорости выполнения запросов.

6. Безопасность:
    - Модель может работать полностью оффлайн, что обеспечивает конфиденциальность данных.
7. Интерфейсы программирования:
    - Поддерживается Python SDK для упрощения интеграции и разработки приложений.
8. Масштабируемость:
    - Может масштабироваться до кластеров машин для обработки большого количества запросов одновременно.
9. Кастомизация:
    - Пользователи могут адаптировать и настраивать параметры модели для своих нужд.
10. Документация и поддержка сообщества:
    - Обширная документация и активное сообщество разработчиков, предоставляющее поддержку и обновления.

vLLM — быстрая и простая в использовании библиотека для обслуживания LLM. Она предназначена для упрощения процесса вывода данных и развертывания моделей на серверах. Основные преимущества vLLM включают:

1. Высокая производительность: Библиотека оптимизирована для работы с большими моделями, обеспечивая высокую скорость обработки запросов.
2. Простота интеграции: vLLM легко интегрируется в существующие системы благодаря простому API.
3. Поддержка различных моделей: поддерживает различные архитектуры нейронных сетей, такие как трансформеры и другие современные модели.
4. Масштабируемость: позволяет масштабировать работу с моделями на несколько серверов или кластеров.

5. Удобство развертывания: обеспечивает простые инструменты для развертывания моделей в производственных средах.
6. Открытый исходный код: это open-source проект, что позволяет разработчикам вносить изменения и улучшения в соответствии со своими потребностями.

vLLM может использоваться в различных приложениях, таких как чат-боты, системы генерации текста, анализ текстов и многое другое.

LangChain — это мощная и гибкая библиотека на Python, предназначенная для построения цепочек (chain) между различными компонентами искусственного интеллекта, такими как большие языковые модели (LLM), базы знаний, API и другие источники данных. Основная цель LangChain — упростить создание сложных приложений на базе искусственного интеллекта, обеспечивая удобный способ комбинирования различных инструментов и технологий.

Основные преимущества vLLM включают:

1. Модульность: предоставляет множество готовых компонентов (agents, chains, memory и др.), которые можно легко комбинировать для решения конкретных задач.
2. Поддержка различных источников данных: возможность подключать разные типы данных, такие как документы, базы данных, API, а также различные языковые модели.
3. Инструменты для обработки естественного языка: включает в себя множество функций для работы с текстом, таких как извлечение информации, обработка вопросов и ответов, генерация контента и многое другое.
4. Агентное программирование: позволяет создавать агентов, которые могут выполнять сложные задачи, используя комбинации различных источников данных и моделей.

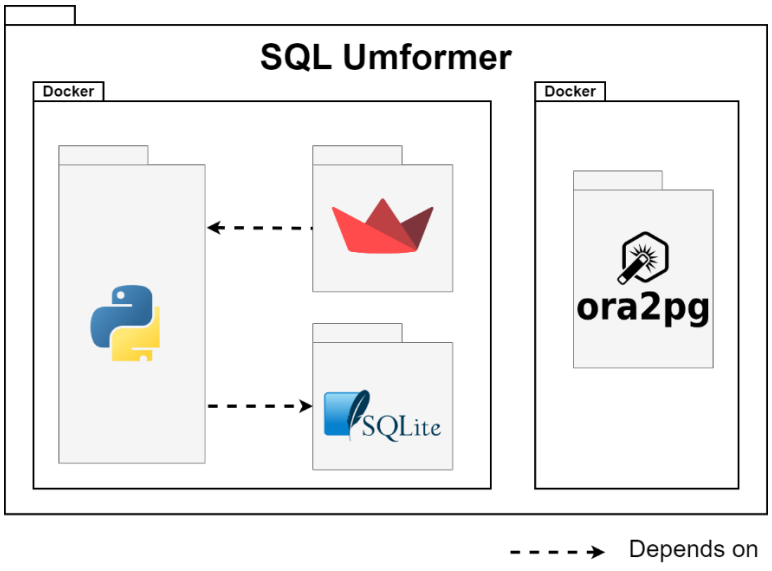
5. Интерактивные приложения: возможность разрабатывать интерактивные чат-боты, ассистенты и другие приложения, взаимодействующие с пользователем.
6. Совместимость с разными платформами: поддержка работы с различными облачными сервисами, такими как OpenAI, Hugging Face, Cohere и другими.

Использование открытых моделей и фреймворков обеспечит прозрачность и возможность модификации ПО.

ПО предоставляет возможность гибкого выбора применяемой LLM-модели и ее параметров. Пользователи могут выбирать модели и настраивать параметры, чтобы оптимизировать процесс преобразования под свои нужды.

ПО ADESU использует контейнеры.

Функционал разделен на независимые контейнеры, каждый из которых прошел проверку безопасности и имеет минимальное количество уязвимостей.



Архитектура ПО ADESU