УТВЕРЖДАЮ
Директор QANYSH.AI Ltd.
_____ Мукашев Б.
«_____» _____ 20__ г.

ТЕХНИЧЕСКОЕ ЗАДАНИЕ

на разработку универсального модуля очистки, нормализации и синхронизации табличных данных и временных рядов

Заказчик: QANYSH.AI Ltd.

Астана, 2026

1 Общие сведения

1.1 Наименование

Универсальный модуль очистки, нормализации и синхронизации табличных данных и временных рядов.

1.2 Заказчик

Заказчик: QANYSH.AI Ltd. Направление: ИКТ, инженерия данных, качество данных, ETL. Контактное лицо и срок приёма откликов указываются при публикации на платформе Astana Hub.

1.3 Основание

Задание опубликовано в разделе «Технологические задачи» платформы Astana Hub. Исполнитель выбирается заказчиком из числа поступивших откликов. Работы выполняются на основании договора заказной разработки.

1.4 Термины и сокращения

Термины и сокращения, используемые в настоящем задании, приведены в таблице 1.

Таблица 1 - Термины и сокращения

№ п/п	Сокращение	Расшифровка
1	ETL	Извлечение, преобразование и загрузка данных
2	CSV	Текстовый формат табличных данных с разделителями
3	XLSX	Формат электронных таблиц
4	Parquet	Колоночный формат хранения табличных данных
5	CLI	Интерфейс командной строки
6	API	Программный интерфейс
7	Временной ряд	Последовательность значений, привязанных к меткам времени
8	M / D	Приоритет требования: M - обязательное, D - желательное

2 Назначение и цели

2.1 Цель работ

Разработать интегрируемый программный модуль для автоматизированной подготовки разнородных данных к аналитике и машинному обучению.

2.2 Результат работ

Программный модуль, который принимает исходные наборы данных, применяет явно заданные правила, формирует нормализованный результат и

проверяемый отчёт о качестве. Исходные данные не изменяются без явного разрешения.

3 Описание проблемы

Данные из учётных, транзакционных, CRM, событийных и производственных систем поступают в несовместимых форматах. Ручное согласование схем и временных меток снижает воспроизводимость аналитики и повышает риск ошибок.

Характерные проблемы:

- различающиеся кодировки и разделители;
- неоднозначные форматы дат и часовые пояса;
- дубли и пропуски;
- смещение столбцов;
- нерегулярные события и агрегаты с различной частотой;
- аномальные значения.

4 Исходные данные и интеграции

Источники данных, форматы и способы доступа приведены в таблице 2.

Таблица 2 - Источники данных и способы доступа

№ п/п	Источник	Формат	Способ доступа
1	Выгрузки из учётных систем и CRM	CSV, XLSX	Файловые выгрузки, передаваемые заказчиком
2	Реляционная база данных	PostgreSQL	Подключение с правами только на чтение к тестовой базе заказчика
3	Аналитические наборы	Parquet	Файловые выгрузки, передаваемые заказчиком
4	Журналы событий с метками времени	CSV, JSON	Файловые выгрузки, передаваемые заказчиком
5	Результат работы модуля	CSV, Parquet, JSON	Запись в каталог или таблицу, указанные в конфигурации

Заказчик до начала работ предоставляет исполнителю тестовые наборы данных (синтетические или обезличенные), описание схем и справочников, правила нормализации, а также реквизиты доступа к тестовой базе PostgreSQL. Ответственным за предоставление данных и документации со стороны заказчика является контактное лицо, указанное в карточке задачи.

Модуль предоставляет интерфейс командной строки и программный интерфейс (Python API) для встраивания в аналитические процессы заказчика. Запуск обработки выполняется по файлу конфигурации, в котором задаются источники, правила и место записи результата.

5 Требования к модулю

5.1 Функциональные требования

Функциональные требования приведены в таблице 3.

Таблица 3 - Функциональные требования

ID	Приоритет	Требование
F-01	M	Импорт CSV, XLSX, Parquet и выгрузок PostgreSQL
F-02	M	Определение кодировки, разделителя, форматов дат; нормализация схемы
F-03	M	Обработка часовых поясов и явное правило для неоднозначных временных меток
F-04	M	Выравнивание временных рядов с настройкой интервала и агрегации: сумма, среднее, последнее
F-05	M	Обнаружение точных дублей, пропусков и аномальных значений
F-06	M	Маркировка или заполнение по выбранному правилу с сохранением исходных значений
F-07	M	Отчёт: пропуски, дубли, аномалии, исправления, статус качества
F-08	D	Журнал преобразований и воспроизводимый запуск по конфигурации
F-09	D	CLI, API или библиотечный интерфейс для встраивания
F-10	D	Экспорт нормализованного результата и отчёта в машиночитаемом виде

5.2 Нефункциональные требования

Нефункциональные требования приведены в таблице 4.

Таблица 4 - Нефункциональные требования

№ п/п	Область	Требование
1	Развёртывание	Локальный запуск без передачи данных во внешнее облако
2	Технологии	Python, Polars, Pandas, Arrow; другой стек допускается по согласованию с заказчиком
3	Масштаб	Обработка от сотен тысяч до нескольких миллионов строк; показатели подтверждаются тестами
4	Прозрачность	Явные правила изменения данных; отсутствие скрытой коррекции
5	Безопасность	Отсутствие телеметрии и внешних вызовов, не предусмотренных заданием
6	Сторонние компоненты	Допускаются компоненты, не ограничивающие коммерческое использование и не требующие раскрытия кода заказчика (например, MIT, BSD, Apache 2.0)

6 Порядок приёмки

6.1 Приёмочные испытания

Приёмка проводится на синтетических наборах данных: (А) события с метками времени в разных форматах и часовых поясах; (Б) почасовые агрегаты; (В) табличная выгрузка с дублями, пропусками и аномалиями. Модуль должен привести данные к общей схеме и временной шкале, сформировать журнал обработки и отчёт о качестве. Контрольные результаты и параметры тестовой среды согласуются сторонами до начала испытаний.

6.2 Критерии приёмки

Критерии приёмки приведены в таблице 5.

Таблица 5 - Критерии приёмки

№ п/п	Критерий	Содержание проверки
1	Полнота	Реализация всех требований с приоритетом M (F-01...F-07)
2	Достоверность	Корректность выравнивания и отчётности на наборах А, Б, В
3	Сохранность данных	Исходные данные не изменяются, все исправления отражены в журнале
4	Воспроизводимость	Повторный запуск с той же конфигурацией даёт тот же результат
5	Документация	Руководство пользователя и описание интерфейсов переданы заказчику

7 Стоимость и порядок оплаты

Стоимость работ составляет 500 000 (пятьсот тысяч) тенге. Оплата производится поэтапно в соответствии с таблицей 6 на основании подписанных актов выполненных работ.

Таблица 6 - Этапы работ и оплаты

№ п/п	Этап	Результат этапа	Оплата
1	Согласование архитектуры и плана работ	Утверждённые архитектура, план и конфигурация модуля	30 %, 150 000 тенге
2	Демонстрация прототипа	Работающий прототип на наборах А, Б, В	40 %, 200 000 тенге
3	Приёмка и передача результатов	Акт приёмки, переданные исходный код и документация	30 %, 150 000 тенге

8 Передача результатов и права

Результаты работ, включая модуль, исходный код и документацию, передаются заказчику с исключительными правами на условиях договора заказной разработки.

В работе используются только синтетические или надлежащим образом обезличенные данные. Персональные, клиентские и коммерчески чувствительные сведения не предоставляются.

9 Состав отклика исполнителя

Отклик должен содержать сведения, приведённые в таблице 7.

Таблица 7 - Состав отклика исполнителя

№ п/п	Что направить	Содержание
1	Предложение	Подход к реализации, архитектура, стек
2	Опыт	Сведения о выполненных аналогичных работах и составе команды
3	Сроки	Календарный план разработки и приёмки
4	Стоимость	Согласие с суммой и этапами оплаты по разделу 7 либо обоснованное предложение
5	Контакт	Ответственное лицо и контактные данные