**Практическая работа: Проектирование процесса обработки данных для аналитической платформы на основе NoSQL**

**Тема:** Моделирование и оптимизация ETL-процесса загрузки данных в колоночную NoSQL-БД

**Цель задания:**

* Освоить принципы проектирования процессов работы с данными в NoSQL-окружении
* Получить навыки моделирования процессов преобразования данных (CSV → JSON → NoSQL)
* Научиться проектировать эффективные конвейеры данных для аналитических задач

**Условие (контекст):**  
Вы — data-инженер в компании, которая разрабатывает аналитическую платформу для интернет-магазина. Ваша задача — спроектировать и оптимизировать процесс регулярной загрузки данных о продажах из различных источников в колоночную NoSQL-БД (ClickHouse) для последующего анализа.

**Исходные данные:**

* Ежедневные CSV-файлы с заказами из разных региональных филиалов
* JSON-логи действий пользователей на сайте
* Требуется обеспечить ежедневное обновление аналитических отчетов

**Задачи:**

**1. Моделирование процесса (BPMN)**  
Постройте BPMN-диаграмму процесса ETL (Extract, Transform, Load), включая:

* Извлечение данных из CSV и JSON-источников
* Преобразование и валидацию данных
* Загрузку в ClickHouse
* Формирование отчетов

**2. Анализ проблем текущего процесса**  
Выявите 5 потенциальных проблем в процессе работы с данными:

* Ручные операции преобразования форматов
* Отсутствие валидации данных
* Дублирование данных
* Низкая производительность загрузки
* Отсутствие мониторинга качества данных

**3. Проектирование оптимизированного процесса**  
Предложите оптимизированный процесс to-be с:

* Автоматизацией преобразования CSV → JSON → ClickHouse
* Добавлением этапов валидации и очистки данных
* Механизмом обработки ошибок и повторной обработки
* Системой мониторинга качества данных

**4. Реализация преобразования данных**  
Напишите псевдокод или опишите логику скрипта для:

* Чтения CSV-файла с заказами
* Преобразования в JSON с нужной структурой
* Валидации данных (проверка форматов, обязательных полей)
* Загрузки в ClickHouse

**Пример структуры JSON:**

json

{

"order\_id": "1001",

"region": "Moscow",

"order\_date": "2023-10-25",

"items": [

{

"product\_id": "301",

"quantity": 2,

"price": 1500.00

}

],

"total\_amount": 3000.00

}

**5. Разработка KPI для процесса**  
Определите 3 ключевых метрики для оценки эффективности процесса:

* Время от получения данных до доступности в отчетах
* Процент успешно обработанных данных
* Объем обрабатываемых данных в день

**Ожидаемые результаты:**

**1. BPMN-диаграмма as-is:**

* Подробная диаграмма текущего процесса с указанием участников
* Выделенные проблемные области

**2. Анализ проблем:**

* 5 конкретных проблем с обоснованием влияния на процесс
* Оценка потерь времени и ресурсов

**3. BPMN-диаграмма to-be:**

* Оптимизированный процесс с автоматизацией ключевых этапов
* Добавленные этапы валидации и контроля качества

**4. Скрипт преобразования данных:**

* Логика обработки и преобразования данных
* Механизм обработки ошибок и валидации

**5. Система KPI:**

* 3 измеримых показателя эффективности процесса
* Методика расчета каждого показателя

**Критерии оценки:**

| **Критерий** | **Баллы** | **Описание** |
| --- | --- | --- |
| **Полнота BPMN-модели** | 0-4 | Детализация процесса, использование элементов BPMN |
| **Глубина анализа проблем** | 0-3 | Конкретность проблем, обоснование критичности |
| **Каство предложений по оптимизации** | 0-3 | Практическая реализуемость, эффективность |
| **Корректность преобразования данных** | 0-2 | Правильность логики преобразования CSV → JSON |
| **Измеримость KPI** | 0-2 | Конкретность метрик, возможность измерения |

**Шкала оценки:**

* **12-10 баллов** — отлично
* **9-7 баллов** — хорошо
* **6-4 балла** — удовлетворительно
* **менее 4 баллов** — неудовлетворительно