В данном кейсе необходимо разработать систему для конвертации PDF документов в текстовые файлы с сохранением структурных элементов. Особое внимание уделяется корректному извлечению и преобразованию таблиц в структурированные таблицы, а также схем и диаграмм в соответствующие форматы (например, в виде изображений или векторных схем). Система предназначена для обработки базы данных документов компании, в которой большинство PDF не содержит текстового слоя, а часть документов имеет текстовый слой.

3.1 Конвертация PDF документов

• Обработка документов с текстовым слоем: извлечение текста, форматирование и сохранение структуры исходного документа.
• Обработка документов без текстового слоя: применение OCR для распознавания текста, извлечение информации и последующее форматирование.
• Корректное преобразование таблиц: извлечение табличных данных с сохранением ячеечной структуры, разбиение на строки и столбцы.
• Извлечение схем и диаграмм: определение графических элементов (схем, диаграмм) и их сохранение в виде изображений или векторных схем с сохранением оригинальной композиции.

3.2 Интеграция с базой данных

• Импорт документов из существующей базы данных компании.
• Обработка различных типов PDF файлов (с текстовым слоем и без него) в едином рабочем процессе.
• Экспорт результатов конвертации в виде текстовых файлов с дополнениями (таблицы, схемы).

3.3 Ручная корректировка и проверка

• Интерфейс для просмотра исходного и конвертированного документа с возможностью ручного редактирования результатов.
• Функционал для корректировки извлечённых таблиц и схем в случае ошибок автоматического распознавания.

№	ФИО	Должность	Отдел	Возраст	Дата приема на работу
1	Иванов Петр А.	Менеджер	Продажи	35	01.06.2018
2	Сидорова Мария В.	Аналитик	IT	29	15.09.2019
3	Кузнецов Сергей Н.	Разработчик	IT	31	10.03.2020
4	Петрова Анна С.	Дизайнер	Маркетинг	27	22.07.2021
5	Смирнов Дмитрий К.	Тестировщик	IT	33	05.11.2017
6	Волкова Елена П.	HR-менеджер	HR	38	12.02.2016
7	Морозов Алексей В.	Маркетолог	Маркетинг	40	18.05.2015
8	Ковалева Ольга Д.	Бухгалтер	Финансы	45	25.08.2014
9	Николаев Игорь С.	Инженер	Производство	50	03.04.2013
10	Федорова Татьяна А.	PR-менеджер	Маркетинг	36	14.12.2018
11	Андреев Максим П.	Юрист	Юридический	42	30.01.2017
12	Васильева Дарья Л.	Секретарь	Администрация	28	19.06.2020
13	Григорьев Илья В.	Логист	Логистика	39	27.03.2016
14	Романова Екатерина	Программист	IT	26	08.09.2022
15	Соколов Артем Н.	Администратор	IT	32	11.07.2018

Проведение анализа затрат на разработку, внедрение, эксплуатацию, поддержку и масштабирование системы.

• Оценка стоимости облачных сервисов, серверного оборудования, лицензий и трудовых ресурсов команды.
• Составление подробной сметы проекта с указанием статей затрат и прогнозируемых сроков окупаемости, документирование результатов оценки.

4. Нефункциональные требования

• Производительность: система должна эффективно обрабатывать большие объёмы документов с использованием параллелизма, кэширования и оптимизированных алгоритмов OCR.
• Масштабируемость: возможность обработки растущего объёма документов и интеграции с новыми источниками данных.
• Надёжность: корректная обработка ошибок, устойчивость к сбоям, обеспечение целостности данных.
• Юзабилити: интуитивно понятный интерфейс для просмотра, редактирования и контроля результатов конвертации, адаптированный для различных устройств.

5. Архитектура решения

5.1 Компоненты