

**Библиотеки подпрограмм (SDK) для текстопонимания и текстогенерации
на основе технологий машинного обучения**

Новизна решения



Онтологический подход - в качестве основополагающей методологии обработки и анализа корпоративной документации, а так же извлечения требуемых данных из них.



Высокая скорость обработки документов за счёт применяемого стека алгоритмов машинного обучения и извлечения информации из корпоративной документации.



Наличие открытых задокументированных интерфейсов интеграции, а также примеров кода интеграции, обеспечит возможность информационного взаимодействия с внешними источниками данных и позволят встраивать библиотеки в корпоративные информационные системы и ресурсы.



Использование собственных методик и правил извлечения данных из корпоративных документов.

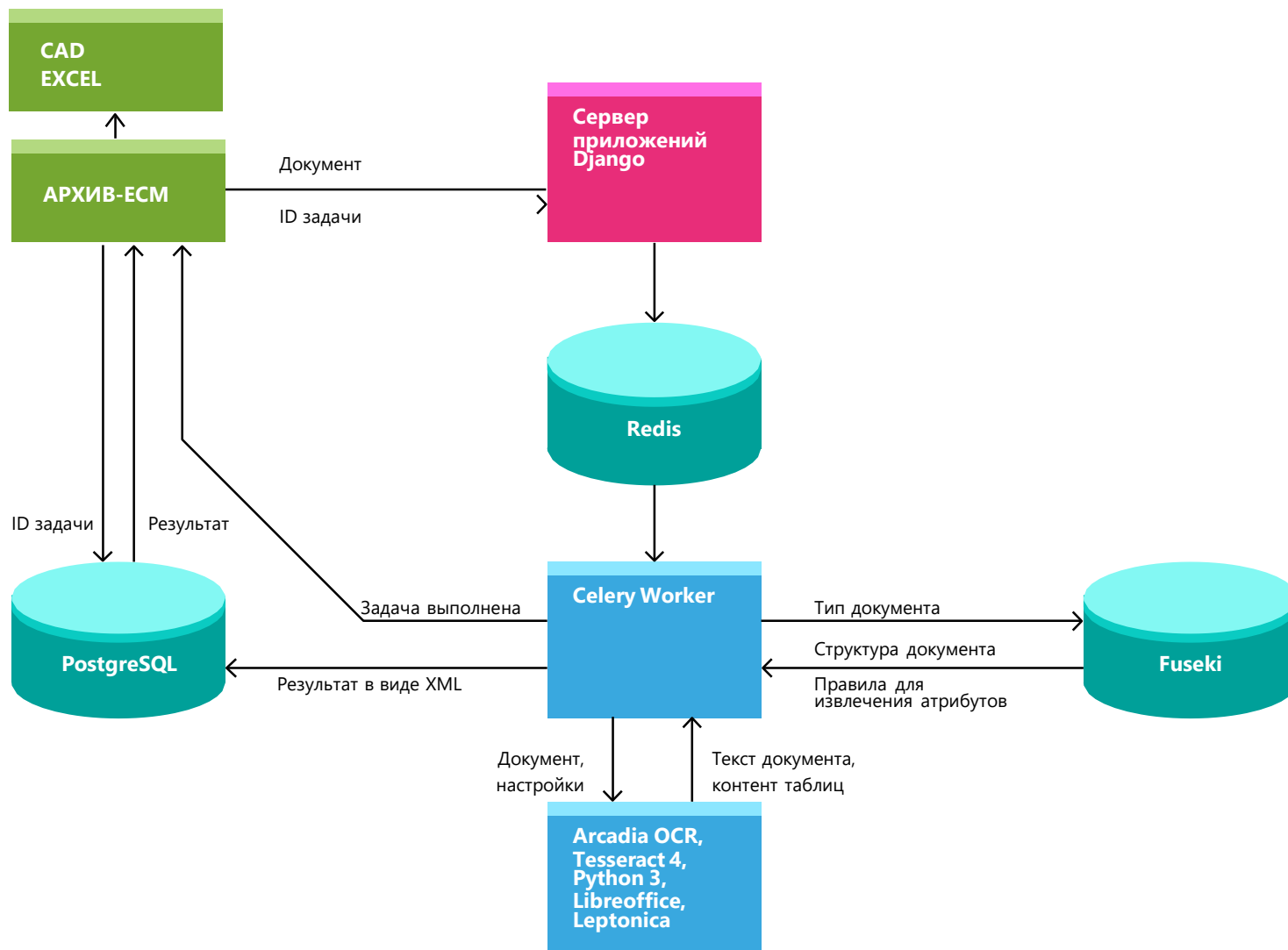
- ■ ■ Библиотеки подпрограмм (SDK) для текстового понимания и текстогенерации на основе технологий машинного обучения

Основа решения – Собственная разработка компании

SDK входит в состав программного комплекса управления цифровым контентом «ЕСМ-Интеллект».

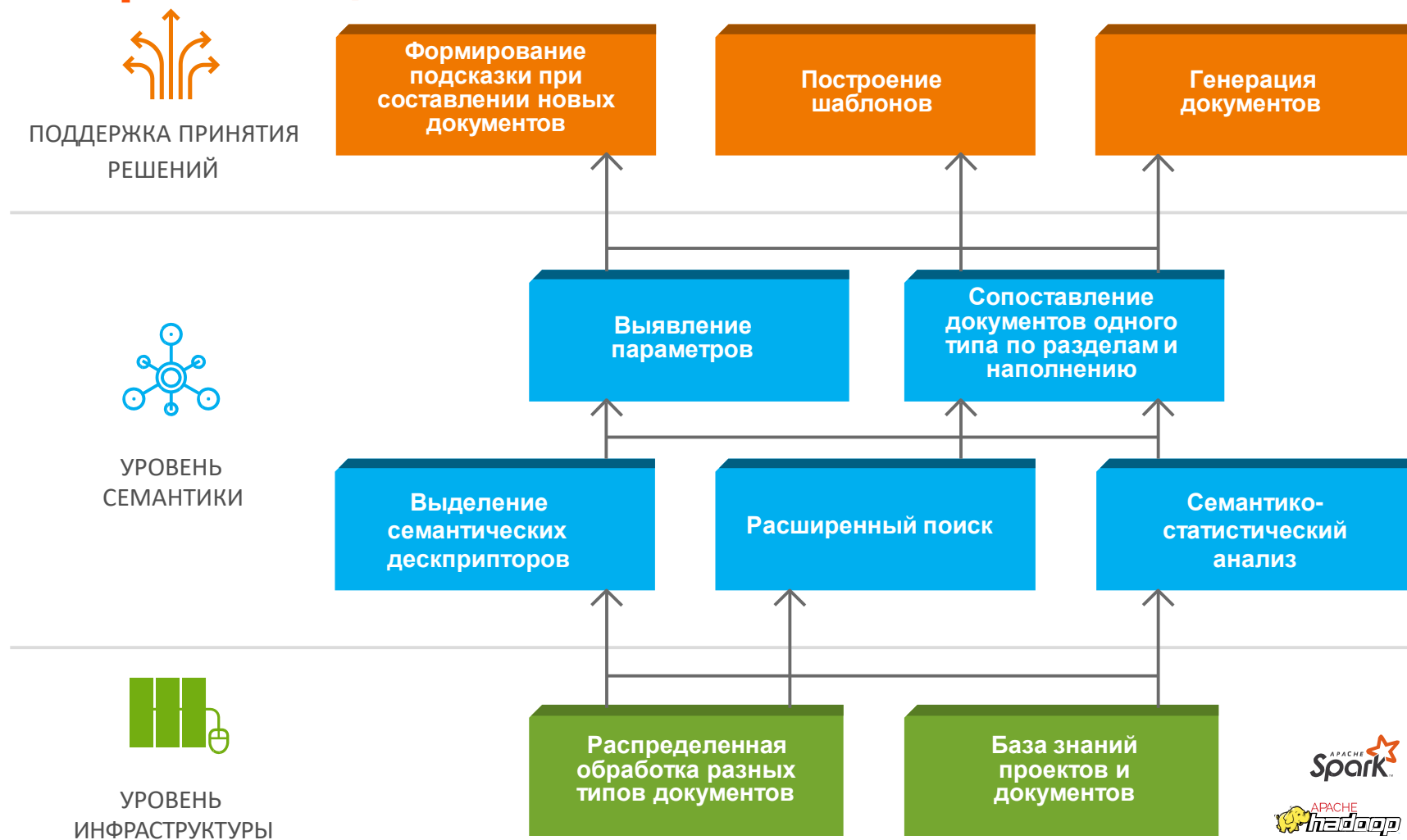


Архитектура решения



- **CAD** – CAD-система, TeamCentr или любая другая система, которую необходимо обеспечить инструментом текстового понимания и текстогенерации
- **PostgreSQL** – хранилище временных файлов
- **Сервер приложений** – принимает запросы, отправляет документы в базу данных и ставит исполнение в очередь задач
- **Redis** – очередь задач
- **Celery worker** – сервер-исполнитель программного кода, выполняет логику работы для поступающих из очереди (редис) задач
- **Arcadia OCR** – распознавание текста, извлечение данных
- **Fuseki** – СУБД для хранения онтологий.

Уровни реализации



Уровни реализации



- ■ ■ Библиотеки подпрограмм (SDK) для текstopонимания и текстогенерации на основе технологий машинного обучения

Семантическая обработка данных

The screenshot displays the 'Автоматизированная система текstopонимания и текстогенерации' (Automated text understanding and generation system) interface. The main window is divided into several panels:

- Документы (Documents):** A sidebar on the left with a search bar and a list of documents. The selected document is 'Стандарт на базе 34.01-6-005-2019'.
- Редактирование (Editing):** The central panel shows the document '34.01-13-016-2020_1.pdf' in edit mode. The title is 'Титульный лист' (Title page). The text includes: 'ПУБЛИЧНОЕ АКЦИОНЕРНОЕ ОБЩЕСТВО «РОССИЙСКИЕ СЕТИ»', 'СТАНДАРТ ОРГАНИЗАЦИИ ПАО «РОССЕТИ»', 'СТО 34.01 -7-202', 'Новый стандарт', 'Стандарт организации', 'Дата введения: 30.09.2021', 'ПАО «Россети»', and 'Предисловие'. The preface text states: 'Цели и принципы стандартизации в Российской Федерации установлены Федеральным законом от 27 декабря 2002 г. № 184-ФЗ «О техническом регулировании», объекты стандартизации и общие положения при разработке и применении стандартов организаций Российской Федерации – ГОСТ Р 1.4-2004 «Стандартизация в Российской Федерации. Стандарты организаций. Общие положения», общие требования к построению, изложению, оформлению, содержанию и обозначению межгосударственных стандартов, правил и рекомендаций по межгосударственной стандартизации и...'.
- Документ в сравнении (Document comparison):** A panel on the right showing a side-by-side comparison of the document with a similar one. It highlights differences in yellow and similarities in pink.
- Похожие документы (Similar documents):** A panel on the far right showing a list of similar documents, including '34.01-13-016-2020_1.pdf'.

Two callout boxes are overlaid on the right side of the interface:

- A green box labeled 'Сходства' (Similarities) points to the pink highlighted areas in the comparison panel.
- A red box labeled 'Различия' (Differences) points to the yellow highlighted areas in the comparison panel.

At the bottom of the editing panel, there are buttons for 'ОТМЕНА' (Cancel) and 'СОХРАНИТЬ' (Save).

Семантический поиск

Основа – базы знаний и онтологии.

Здесь релевантность документа запросу определяется семантически, а не синтаксически. В его основе – более точная интерпретация поисковых намерений пользователя.

Это явление многофакторное, обрабатывающее сложные логически обоснованные запросы, которые невозможно решить за счет традиционных видов поиска информации.



- ■ ■ Библиотеки подпрограмм (SDK) для текстовопонимания и текстогенерации на основе технологий машинного обучения

«Семантический Анализ документов». Пример

