Разработка средств секвенирования файлов описания химических компонентов и системы поиска на основе полученных структур данных

Подготовил:
Рассказов Сергей Михайлович КЭ-222

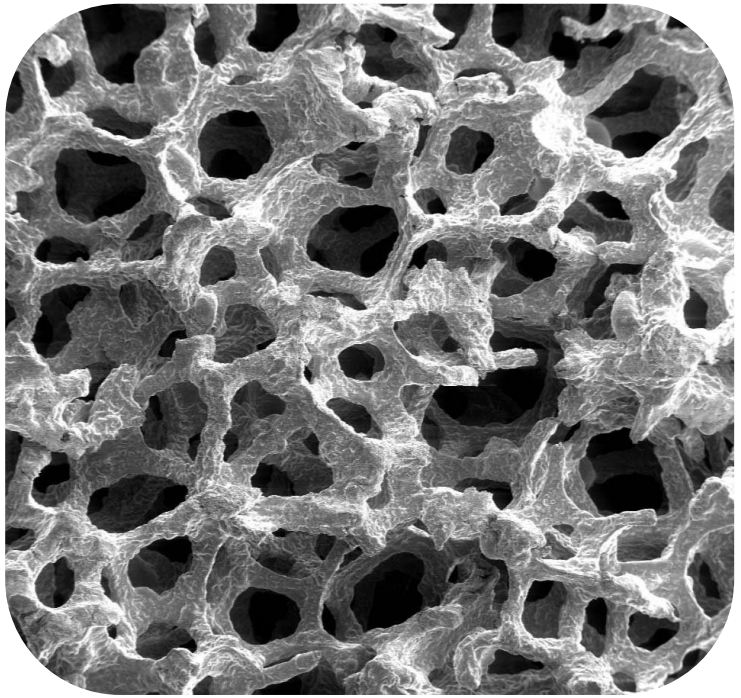
Научный руководитель:
к.т.н. Кафтанников Игорь Леопольдович

Digital Twins for Materials

Технологический институт Джорджии; Surya R. Kalidindi, Michael Buzzy, Brad L. Boyce, Remi Dingreville 16.03.22г.

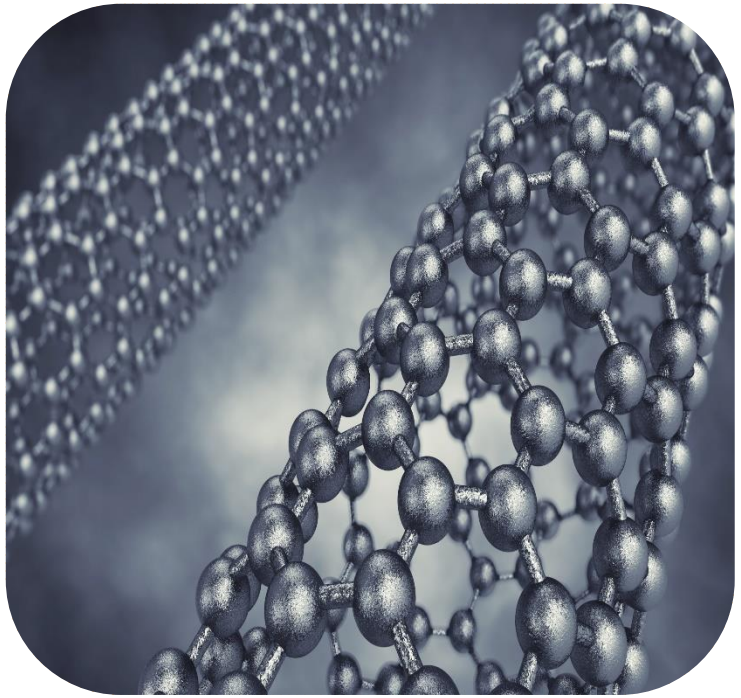
МАКРОУРОВЕНЬ

Физические объекты в метриках метровых, сантиметровых единиц



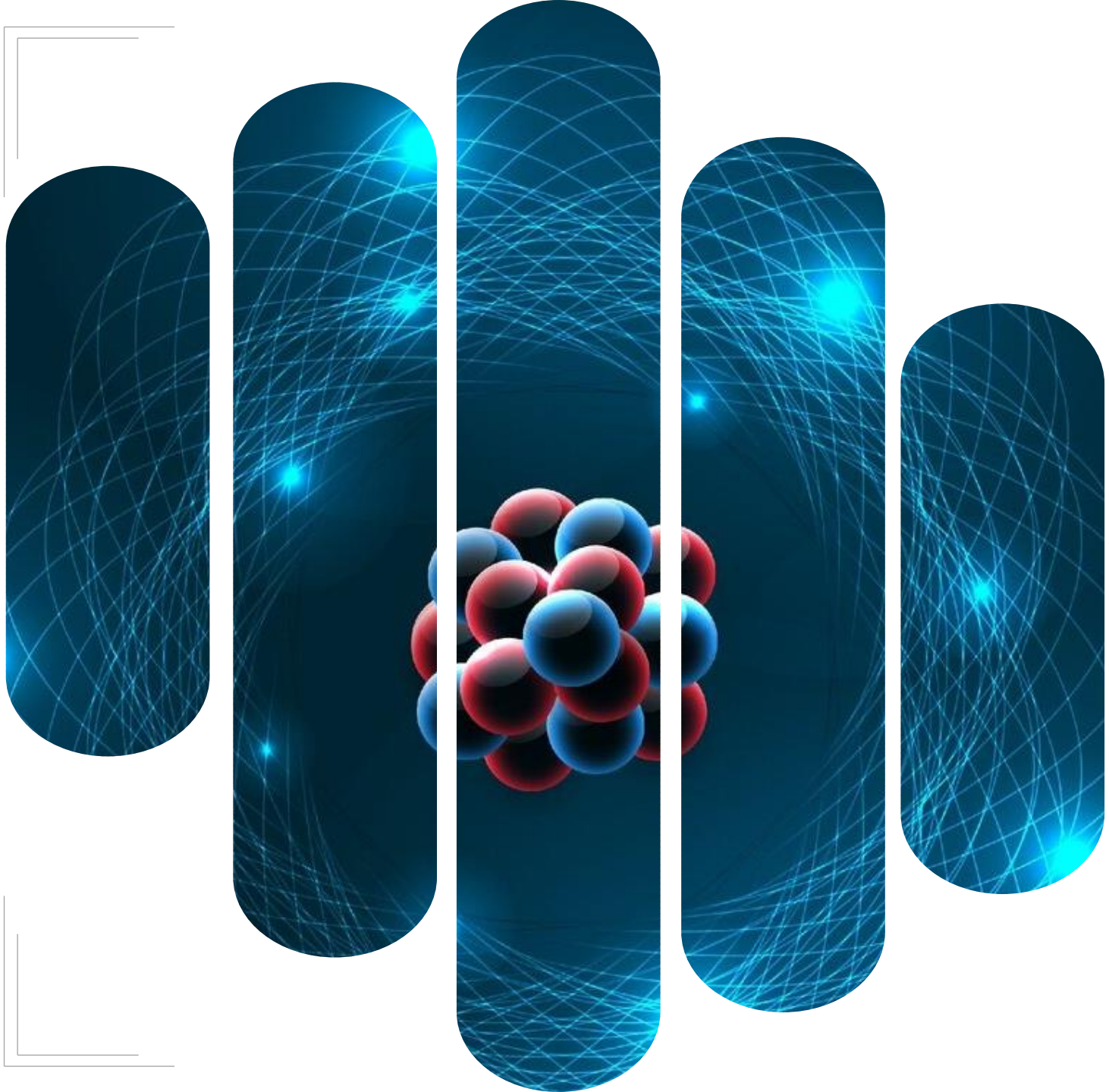
МИКРОУРОВЕНЬ

Сплавы, твердые растворы, органические соединения, агрегации молекул и т.п.



НАНОУРОВЕНЬ

В основном, структуры с нанометрикой, молекулы



АТОМНЫЙ УРОВЕНЬ

Допустимо опираться на эти три уровня, если рассматриваются довольно большие объекты, так в этой статье речь идёт о газотурбинных двигателях. В случае, если речь идёт о цифровых двойниках материалов, то необходимо опускаться на более низкий уровень – атомный. При такой детализации определяются и исследуются химические связи, и силы этих связей. Свойства взаимного расположения атомов

ЦЕЛЬ

Создание платформы, которая в качестве входных данных принимает файлы специфичных файлов содержащих химическую информацию. Выделяет наиболее полезные ее части и сохраняет в базе данных для более удобного представления и ускорения процедур поиска.

Задачи

01

Обзор аналогов

02

Форматы файлов для хранения информации о структурах и свойствах материалов;

03

Определение требований

04

Проектирование Информационной системы

05

Разработка

Аналоги

Имя проекта	Основной тип данных	Доступ	Данные	Архитектура	Основное преимущества	Основной недостаток
Materials Project	Кристаллические материалы	Открытый доступ	Расчетные и экспериментальные	Микросервисная архитектура (WEB интерфейс, API, База данных)	Помощь в подборе методов расчётов, огромный инструментарий	Возможны проблемы с работой из-за большой загруженности
NOMAD	Неорганические соединения	Открытый доступ	Расчетные и экспериментальные	API есть, распределенная БД с поиском Apache Solr	алгоритмы машинного обучения	Модули аналитики на внешних серверах
ICSD	кристаллических структурах неорганических соединений	Подписка\ лицензия	экспериментальные	Закрыто (коммерция)	Довольно полная картина материалов	Доступ
CSD	Органические и неорганические соединения, структуры кристаллов	Лицензия, ограниченная функциональность	экспериментальные	Закрыто (коммерция)	Авторитетность Кембриджский банк структурных данных	Доступ

CIF (Crystallographic Information File)

CIF является последовательным файлом содержащим текстовую информацию в кодировке ASCII, длину строк не более 80 символов и другие технические ограничения. Он формируется по заранее заданным правилам, которые сформулированы и приняты International Union of Crystallography (IUCr) в 1992 году. И описаны в «A Guide to CIF for Authors.»

В 2014 году приняли стандарт CIF2, в котом добавили возможность использовать Unicode и другие

```
_exptl_crystal_description          plate
_exptl_crystal_colour               colourless
_exptl_crystal_size_max             0.30
_exptl_crystal_size_mid             0.28
_exptl_crystal_size_min             0.10
_exptl_crystal_density_meas         ?
_exptl_crystal_density_diffrn       1.210
_exptl_crystal_density_method       'not measured
_exptl_crystal_F_000                266
_exptl_absorpt_coefficient_mu       0.089
_exptl_absorpt_correction_type      Multi-scan
_exptl_absorpt_process_details      '(DENZO-SMN;
Otwinowski & Minor, 1997)'
_exptl_absorpt_correction_T_min     0.950
_exptl_absorpt_correction_T_max     0.988

_diffrn_ambient_temperature         150(1)
_diffrn_radiation_wavelength        0.71073
_diffrn_radiation_type              MoK\a
_diffrn_radiation_source             'fine-focus
sealed X-ray tube'
_diffrn_radiation_monochromator     graphite
_diffrn_measurement_device_type          'Nonius
KappaCCD'
_diffrn_measurement_method            '\f scans,
_diffrn_standards_decay_%           0
_diffrn_reflns_number               8456
_diffrn_reflns_av_R_equivalents     0.064
_diffrn_reflns_av_sigmaI/netI       0.0848
_diffrn_reflns_limit_h_min          -11
_diffrn_reflns_limit_h_max          11  7
```

```
ATOMS IN THE ASYMMETRIC UNIT    5 - ATOMS IN THE UNIT CELL:16
     ATOM                X/A                 Y/B
 ************************************************************
      1 T   6 C     0.000000000000E+00  5.000000000000E-01
      2 F   6 C    -5.000000000000E-01  0.00
      3 T   8 O     0.000000000000E+00  5.00
      4 F   8 O    -5.000000000000E-01  0.00
      5 T   7 N     1.460055586849E-01 -3.53
      6 F   7 N    -1.460055586849E-01  3.53
      7 F   7 N    -3.539944413151E-01 -1.46
      8 F   7 N     3.539944413151E-01  1.46
      9 T   1 H     2.588485243130E-01 -2.41
     10 F   1 H    -2.588485243130E-01  2.41
     11 F   1 H    -2.411514756870E-01 -2.58
     12 F   1 H     2.411514756870E-01  2.58
     13 T   1 H     1.436313742521E-01 -3.56
     14 F   1 H    -1.436313742521E-01  3.56
     15 F   1 H    -3.563686257479E-01 -1.43
     16 F   1 H     3.563686257479E-01  1.43

T = ATOM BELONGING TO THE ASYMMETRIC UNIT
INFORMATION **** fort.34 **** GEOMETRY OUTP

DIRECT LATTICE VECTORS CARTESIAN COMPONENTS
          X                   Y
  0.556500000000E+01   0.000000000000E+00
  0.000000000000E+00   0.556500000000E+01
  0.000000000000E+00   0.000000000000E+00
CARTESIAN COORDINATES - PRIMITIVE CELL
```

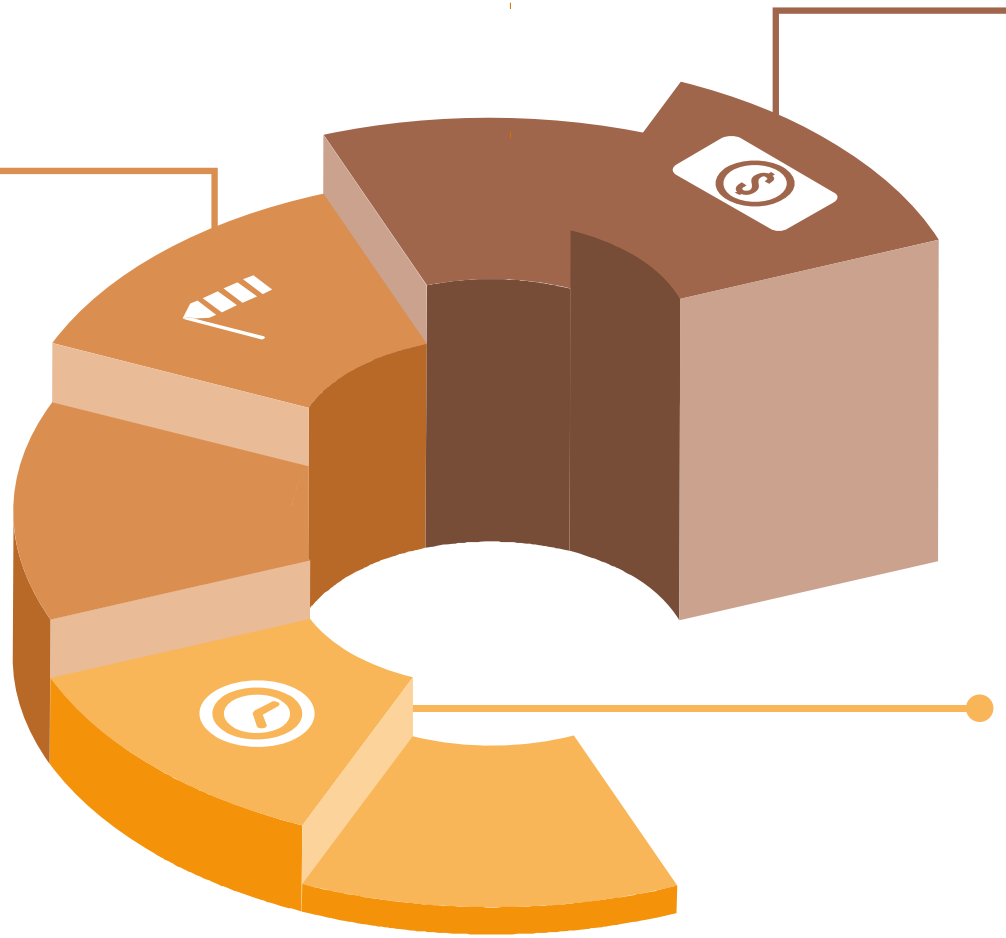
OUT

Является результатом работы программных пакетах по расчету электронной структуры материалов. Формат стал популярным из-за постоянного развития в области химии и материалов. Старые форматы уже не могут покрывать все требования современных исследований. Основным отличием от CIF файла является отсутствие правил по формированию данных, каждая программа делает это по своему, хотя общие черты конечно прослеживаются.

Парсинг

Парсинг

Парсинг – это процесс автоматизированного сбора и структурирование информации из источника при помощи программы или сервиса, для дальнейшей работы с ней, как с отдельными объектами.



Результат

Результатом парсинга является структурированные данные, которые были извлечены из исходного источника информации. Эти данные могут быть представлены в различных форматах, в зависимости от того, какой тип информации вы извлекаете и какой инструмент для парсинга используете.

Источник данных

Источником парсинга может быть любой источник данных, содержащий информацию, которую вы хотите извлечь. Веб-страницы, базы данных, файлы CSV или JSON, XML-документы, файлы логов и др.

Группы

default_group

01

Данные о дате создания документа

submission _details

02

Данные об авторах

processing_ summary

03

Данные о журнале, в котором опубликовано исследование

title_and_ author

04

Понятное название и аннотация

text

05

Данные к рисункам

chemical_data

06

Брутто формула, данные о симметрии, параметрах кристалла и т.д.

refinement _data

07

Уточняющие данные, схема весов, параметры матриц, коэффициенты

atomic _coordinates

08

Данные об атомах, их расположение и т.д.

molecular _geometry

09

Дополнительная информация, не входящая в основной перечень

unknown_ group

10

РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТОВ

Для примера был проведен эксперимент, который показал превосходство разделенной информации по сравнению с последовательной. Данные хранились в локально расположенной базе данных PostgreSQL, на твердотельном накопителе.

CIF: Поиск осуществлялся в группе «Chemical Data», по брутто формуле: «C34 H22 N4 O1 S1».

Количество элементов в базе (шт.)	Время поиска в полных файлах (сек)	Время поиска в разделенных файлах	Во сколько раз быстрее
433	0.011000	0.001000	11,0
4753	0.101000	0.004430	22,8

РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТОВ

OUT оптимизации: Поиск осуществлялся по параметру «band gap», по значению: «4,3846».

Количество элементов в базе (шт.)	Время поиска в полных файлах (сек)	Время поиска в разделенных файлах	Во сколько раз быстрее
404	0.010998	0.000998	11,0
4004	0.091005	0.002002	45,5

Типы OUT файлов

OPTimization

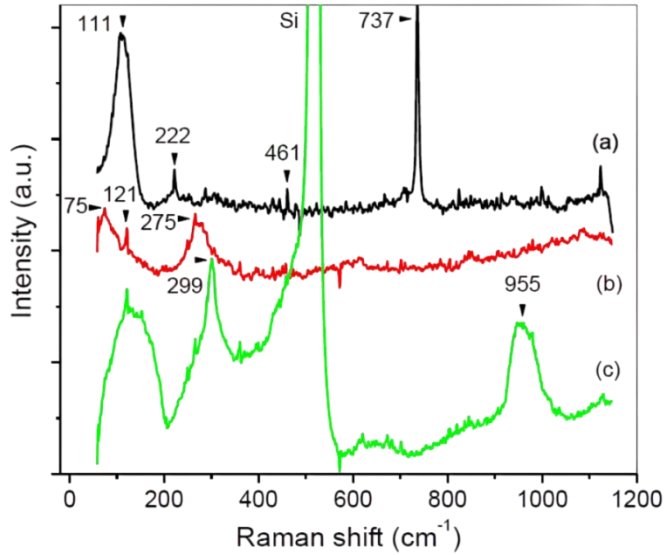
Много итерационный расчёт при котором изменяется положение атомов друг относительно друга с целью выявления самого низкого значения энергии

Other

Другие расчеты, которые занимают около 5% от всех расчётов лаборатории

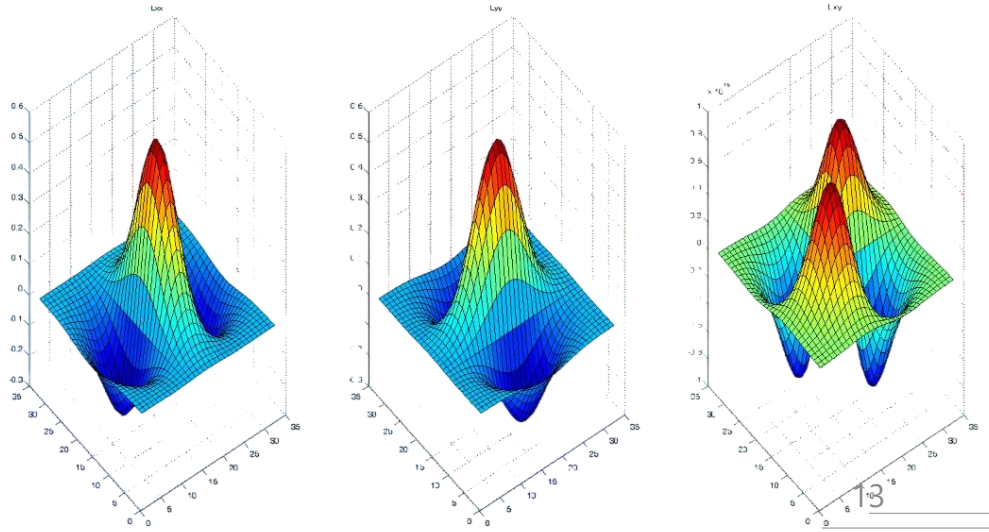
RAMAN спектр

Определение колебательных мод молекул и вибрационных мод в твёрдых телах, который также служит для определения вращательных и других низкочастотных мод систем.



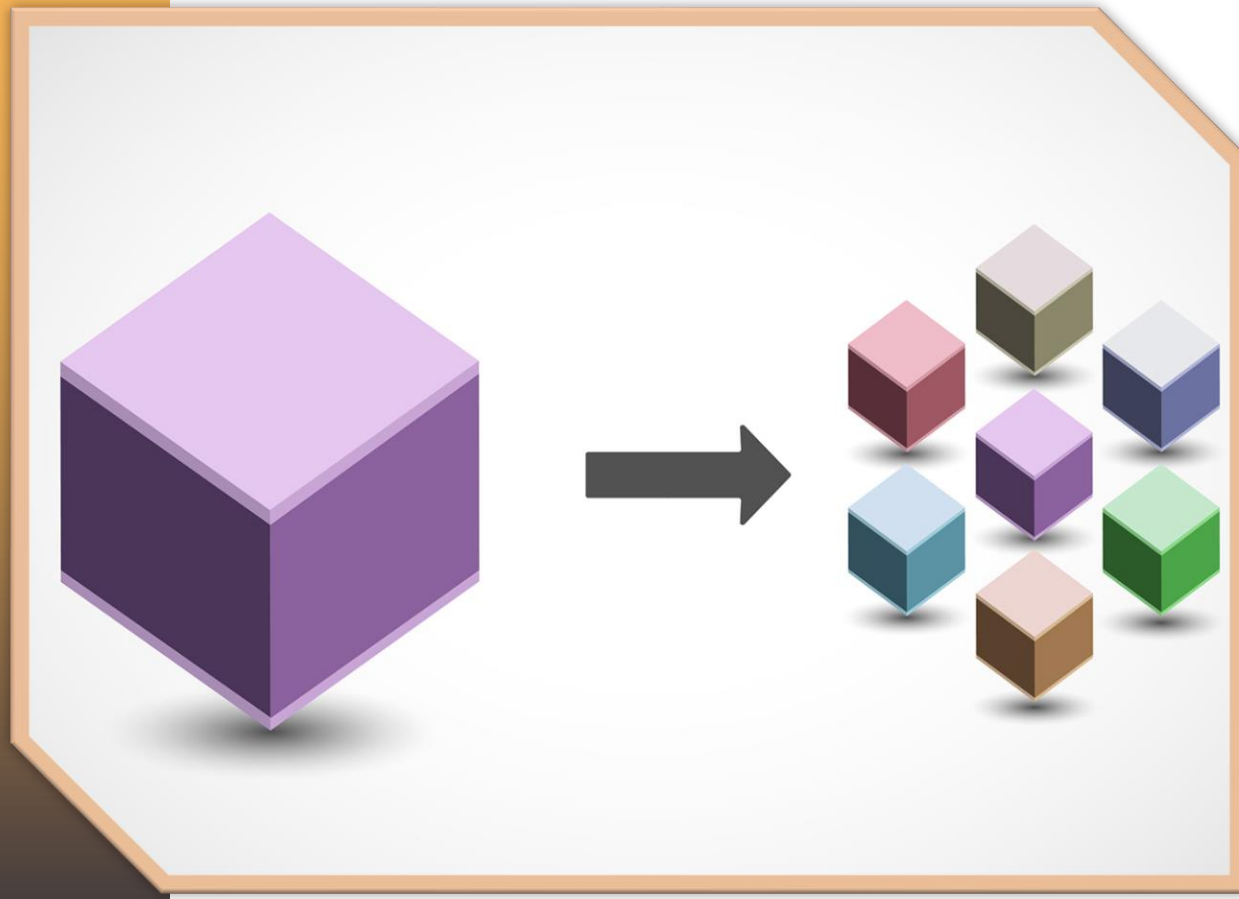
HESSian матрица

По своей сути данный тип расчетов служит для подтверждения данных оптимизации. На сколько, найденное значение в самом деле является минимальным



Микросервисная

Архитектура состоит из изолированных компактных микросервисов, способных расширяться независимо друг от друга.

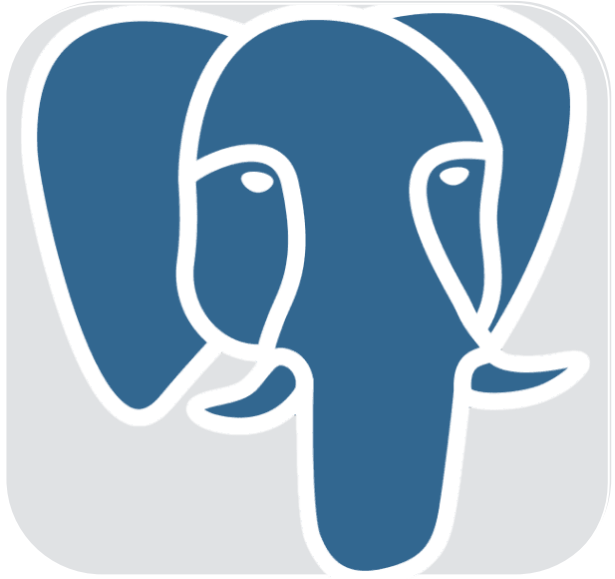


Преимущества

- Предлагает слабую связанность благодаря высокой степени изоляции.
- Повышает модульность.
- Сбой в одном сервисе не затронет всю систему, поскольку они изолированы.
- Предлагает высокую гибкость и масштабируемость.
- Простота модификации может ускорить итерации.
- Позволяет реализовать улучшенную систему обработки ошибок.

Контейнеры

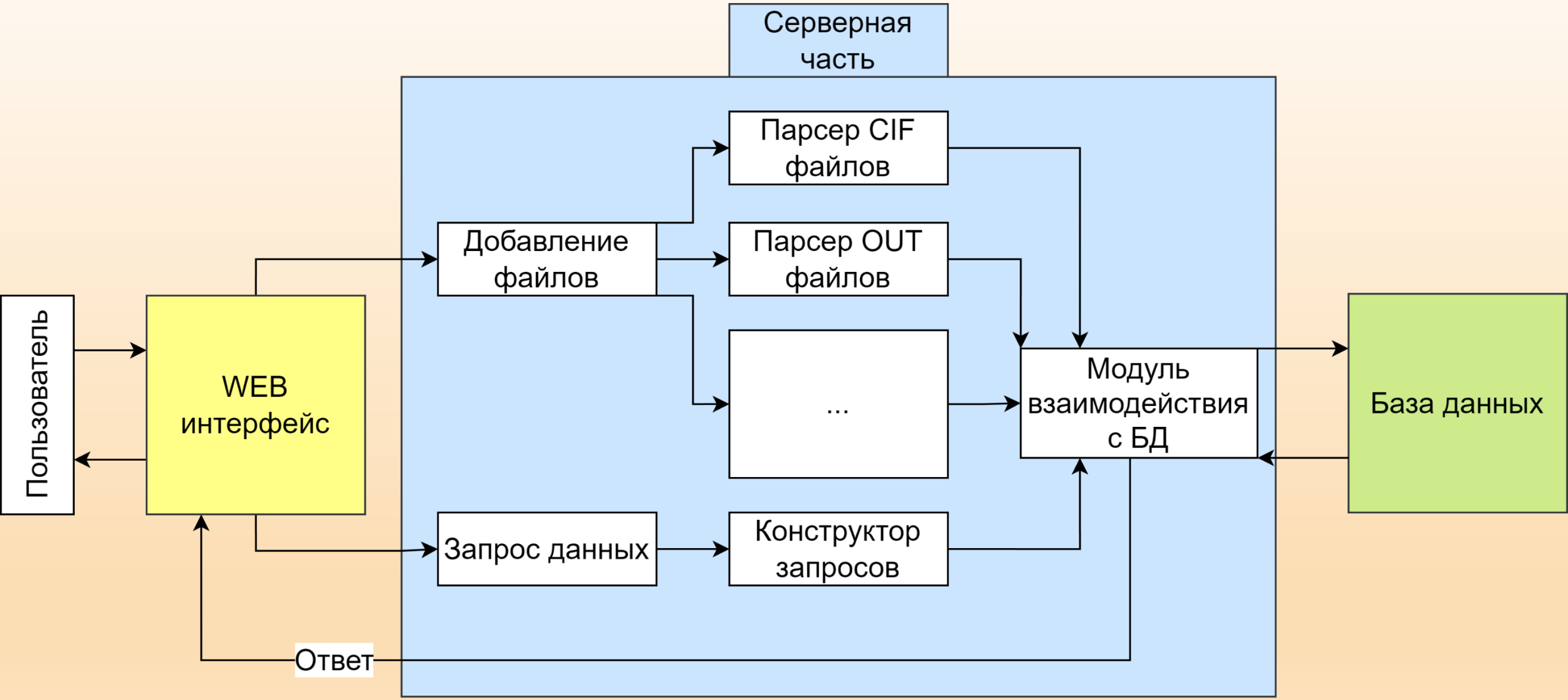
Python Flask



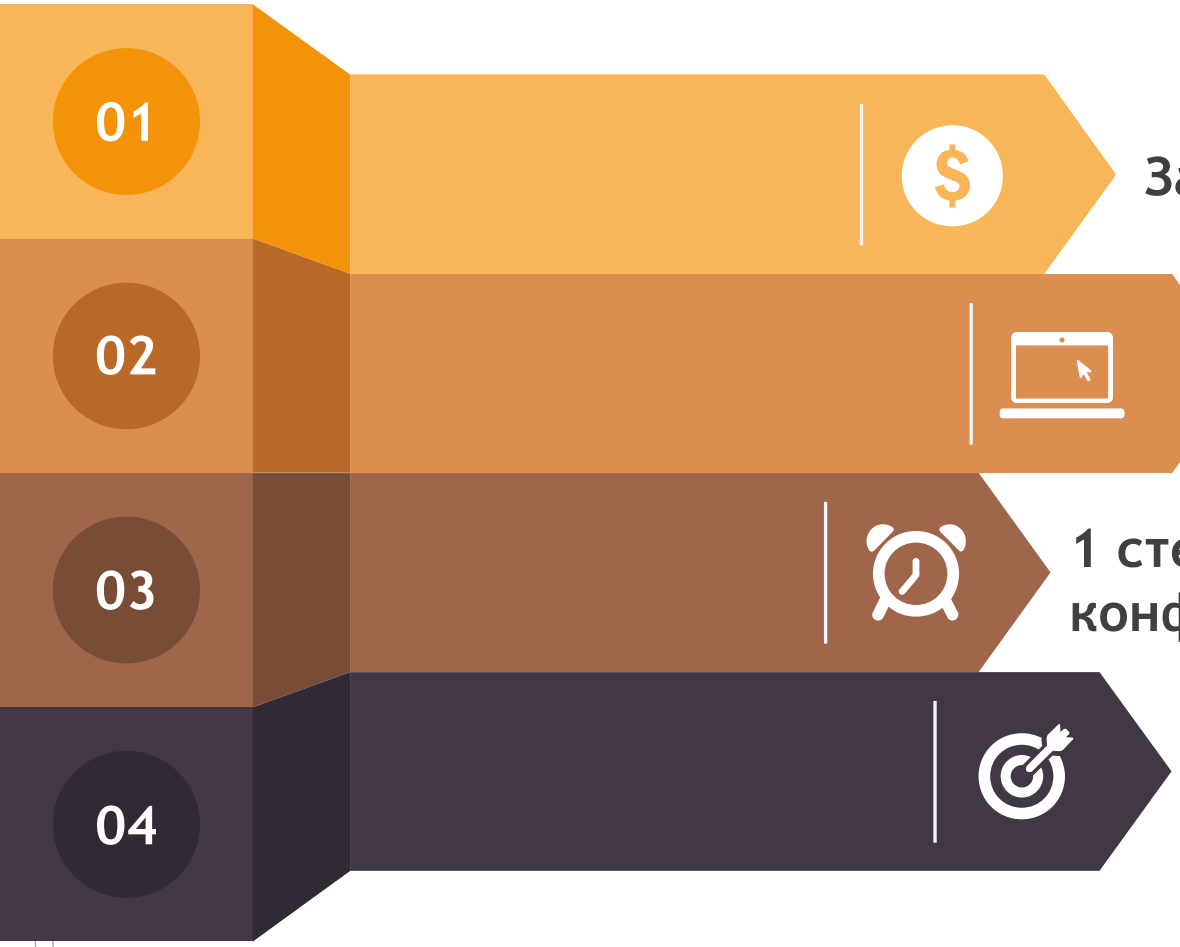
PostgreSQL

Vue

Nginx


Серверная часть
Парсер CIF файлов
Добавление файлов
Парсер OUT файлов
...
Модуль взаимодействия с БД
База данных
Пользователь
WEB интерфейс
Запрос данных
Конструктор запросов
Ответ

Значимые достижения



Заявка на РИД

Статья в сборник "Молодой исследователь. Материалы 10-ой научной выставки-конференции научно-технических и творческих работ студентов".

1 степень за доклад и 3 за стенд на выставке-конференции ЮУрГУ

Проделанная работа включена в отчет по выполнения гранта ЦДМ. Работа одобрена и продлена

Работа описанная в докладе осуществляется в рамках проекта

проект "Цифровой двойник химических соединений и материалов" в рамках СП№2 программы "Приоритет 2030"

Лабораторией Многомасштабного моделирования многокомпонентных функциональных материалов

Под руководством доктора химических наук, доцента, Барташевич Екатерины Владимировны

Спасибо за внимание!

Оглавление

№	Функционал	%
1	Интерфейс пользователя	90%
2	Загрузка файлов CIF, OUT	100%
3	Обработка CIF файлов	100%
4	Хранение и одиночный поиск по CIF данным	100%
5	Многопараметрический поиск по CIF данным	10%
6	Обработка OUT файлов содержащих расчет оптимизации «Crystal 17»	100%
7	Обработка OUT файлов содержащих расчета матрицы Гесса «Crystal 17»	70%
8	Обработка OUT файлов содержащих расчет Рамановского спектра «Crystal 17»	50%
9	Хранение и одиночный поиск по OUT оптимизации	100%
10	Хранение и одиночный поиск по OUT расчета матрицы Гесса	50%
11	Хранение и одиночный поиск по OUT расчета Рамановского спектра	50%
12	Генерация SMILES кодов на основе CIF файлов	90%
13	Анализ модификаций полученных смайлс кодов на основе	5%
13	Объединение полученных данных в структуры по материалу	10%
14	Готовность проекта под серверное размещение	100%
15	Алгоритмы ускорения базы данных (индексирование и кэширование)	10%
16	Адаптация к Hadoop Distributed File System	70%