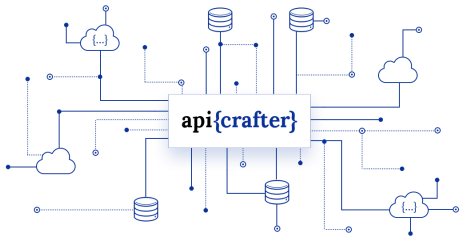


APICrafter.DataCrafter

Инфраструктура организации работы с данными

Данные – новая нефть?

Да, если у вас есть инфраструктура…

- Очистка и хранение данных
- Консолидация данных
- Единый интерфейс доступа для исследований/обработки

… и уверенность в надежности источников

Суровая реальность работы с данными:

Недоступность данных,
Низкое качество данных,
Фрагментированность по различным источникам,
«Зоопарк» систем и программных интерфейсов

Каталоги данных

Какие каталоги данных бывают

Каталоги данных

Порталы открытых данных

- Data.gov
- Hubofdata.ru

Научные репозитории

- Harvard DataVerse

Корпоративные каталоги

- Управление НСИ
  - nsi.rosminzdrav.ru
- Каталоги данных для машинного обучения
  - kaggle.com/datasets

Отраслевые каталоги

- ЕСИМО

Ситуация с классическими каталогами [мета]данных

- Узкозаточены под конкретные задачи
- Не умеют (или очень дорого умеют) идентифицировать типы полей
- Плохо поддерживают не-табличные данные
- Не умеют или умеют дорого автоматически анализировать качество данных

Современные публичные каталоги данных

Data.world

- Большой акцент на визуализации
- Удобное рабочее место
- Пытаются стать Github'ом для данных
- Формируют сообщество

doltHub

- Git-like база данных
- Пытаются стать Github'ом для данных
- Формируют сообщества

Bit.IO

- Делитесь не датасетами, базами данных (с)
- Позволяют сразу создавать свои онлайн СУБД и делиться ими с другими

Какой каталог хотелось бы сделать

1. **Универсальный публичный каталог данных**
2. **Для коммерческого и некоммерческого использования**
3. **Автоматизировать всё что можно автоматизировать**
4. **Научиться работать с данными любого качества**
5. **Собрать много, очень много, данных**
6. **И, конечно, запихнуть туда кучу наработок по работе с данными**

ЕГРЮЛ

Росстат

ФНС России

Собственные базы

... и другие источники

Построение отчетов

Аналитические системы

Web- или мобильные интерфейсы

Интеграция с помощью API

Источники данных | **DataCrafter** | **Система-клиент**

Хранилище

Сбор данных

Очистка данных

Предварительная обработка данных

Что делает DataCrafter?

Сбор данных

Сейчас доступны более чем 200 крупных источников (базы государственных и муниципальных закупок и контрактов, ЕМИСС, Центральный банк, ФНС России, реестры среднего и малого бизнеса, Рособрнадзора и десятки других)

Преобразования в удобные форматы

Накоплен большой опыт по очистке и верификации данных, в том числе в автоматизированном режиме

Актуализация

Регулярность обновления данных в зависимости от набора данных – от 1 дня до 1 месяца. Например, данные о госзакупках обновляются ежедневно

Единая среда

Доступ ко всем данным предоставляется через удобные API и слепки данных

Внешний портал

Возможность поработать с данными до заключения договора на портале beta.apicrafter.ru

Для чего использовать DataCrafter?

Позволяет создать универсальную инфраструктуру данных для различных применений

Обеспечивает возможности интеграции с любыми аналитическими системами

Минимизирует проблемы обеспечения качества и предварительной подготовки данных

Озера данных

Инфраструктура для лабораторий данных

Доступ к данным полу-фабрикатам

Порталы (открытых) данных

Обеспечение данных для работы ситуационных центров

Преимущества

- **Возможность сконцентрироваться на собственных задачах**
- **Консультации и поддержка**
- **Гибкость конфигурирования источников под заказчика**
- **Гарантия доступа к данным**
- **Регулярно обновляемые структурированные данные, прошедшие очистку**
- **Поддержка и постоянное обновление баз данных**

Стабильность и предсказуемость работы – непрерывность ваших бизнес-процессов

Способы доступа к данным

API

Доступ через стандартизованное REST API с поиском записей и получение списка записей по запросу

Сборки данных

Выгрузка полной копии данных в виде базы данных для MongoDB или резервной копии данных в формате JSON lines

Доступ к СУБД

Прямой доступ к базам данных при развертывании на выделенном сервере или в инфраструктуре заказчика

Интерфейс под заказ

Разработка интерфейсов выгрузки в форматах CSV/XML/JSON или веб интерфейсов под задачи заказчика

Статистика по некоторым направлениям

Направление	Всего пакетов	Число записей	Общий объём, Гигабайт
Госфинансы	10*	6,400,000	13
Документы	5	123,000,000	12
Классификаторы	9	240,000	1.4
Люди	8	69,000	0.05
Медицина	9	340,000	0.65
Организации	122	367,000,000	60
Статистика	6	940,000	3
Финансы	6	100,000	0.01

* **Без учёта данных с портала zakupki.gov.ru**

Примеры данных

- Государственные и муниципальные контракты с 2007 года
- Базбы официально опубликованных нормативно-правовых документов
- Государственные и муниципальные закупки с 2013 года
- Протоколы по государственным и муниципальным закупкам с 2013 года
- Закупки малого объёма по субъектам федерации
- Документы стратегического планирования с портала ГАС Управление
- Реестры поставщиков и заказчиков
- ЕГРЮЛ
- Статрегистр юр. лиц
- Бухгалтерские балансы юр. лиц
- Бухгалтерские балансы госучреждений
- Индикаторы статистики Росстата (ЕМИСС)
- Ведомственная статистика ФНС, Минкультуры, Генпрокуратуры и тд.

Запланированные данные

Электронный бюджет

Около 150 пакетов данных, около 150 ГБ.

Данные по госрасходам, госдоходам, участникам бюджетного процесса и так далее.

Порталы открытых данных

Около 15000 наборов данных, около 1000 ГБ.

Данные из региональных порталов открытых данных

Реестры организаций

Сотни реестров организаций таких как торговые реестры, реестры поставщиков, реестры лицензий

Около 500 наборов данных, около 100 ГБ

Особенно значимые данные

Госкаталог культурного фонда, ФИАС, ведомственная статистика, муниципальная статистика

Общий объём около 500 ГБ, около 10 баз данных

Базы нормативных документов

Базы данных НПА федерального уровня и субъектов федерации

Более 1 миллиона документов, более 3 терабайт данных

Поставка

Стоимость инсталляции и интеграции с системами заказчика определяется квалифицированными экспертами после уточнения задач и потребностей

1 **DataCrafter как дистанционный сервис**

Все данные и код хранятся на серверах Поставщика, доступ к системе осуществляется через интернет.

2 **DataCrafter на выделенном сервере**

Мы разворачиваем и управляем выделенным сервером с экземпляром DataCrafter'а в облаке.

3 **DataCrafter на инфраструктуре Заказчика**

Система развертывается на серверах Заказчика, копия данных находится у заказчика и регулярно обновляется с централизованного сервера.

4 **DataCrafter как специальное решение**

Под особые задачи заказчика мы разворачиваем DataCrafter + инфраструктуру для работы исследователей

Необходимые набор компонентов, стоимость инсталляции и интеграции с системами заказчика определяется квалифицированными экспертами после уточнения задач и потребностей Заказчика

Технические подробности

Объёмы данных (Data Crafter)

- **220 коллекций в 180 пакетах данных**

  По организациям, медицине, госфинансам, финансам, медицине и тд.

- **83.3 ГБ структурированных данных**

  Данные подготовленые для доступа через API и сборок данных.

Объёмы данных (специализированные базы)

- **Крупные базы данных**

  ЕГРЮЛ, база госзакупок и госконтрактов, база сведений о госучреждениях

- **Около 1 ТБ структурированных данных**

  Доступны через специализированные API

- **Проходят специальную подготовку**

  Оптимизация под производительность, загрузка в PostgreSQL и разработка специальных API

Объёмы данных (в ожидании к загрузке)

- **200+ каталогов данных и тысячи наборов данных**

  Автоматическая загрузка данных и крупнейших каталогов

- **Около 5 ТБ структурированных данных**

  Базы документов, адресные реестры и иные материалы

- **Проходят оценку востребованности**

  Не все из этих данных реально нужны в регулярной, повседневной работе.

mossmallcontr2016	Контракты малого объёма г. Москвы до 2016 года	Контракты малого объёма г. Москвы до 2016 года включительно	1	574006	616.3MiB	2020-05-08 21:31:04	381
reestrgk	Реестр государственных и муниципальных контрактов за 2007-2011 годы	Реестр государственных и муниципальных контрактов за 2007-2011 годы виде первичных данных	3	2227244	6.6GiB		None
regtaxpassports	Налоговые паспорта субъектов федерации	Налоговые паспорта субъектов федерации с аналитического сайта ФНС России	1	469	107.5KiB	2021-04-22 13:18:41.022000	32
zakmoscontracts	Государственные контракты системы госзаказа города Москвы	Государственные контракты системы госзаказа города Москвы с портала поставщиков города Москвы (zakupki.mos.ru)	1	1722100	1.4GiB	2021-05-01 15:49:10.811000	23

Документы

Набор данных	Название	Описание	Всего таблиц	Всего записей	Объём	Обновлено	Дней назад
fassol	Реестр решений Федеральной антимонопольной службы	Документы решений Федеральной антимонопольной службы	1	517844	165.2MiB	2019-08-31 13:56:10	632

Классификация данных

- **Автоматическое определение типов идентификаторов**

  ИНН, ОГРН, кадастровый код и др. Всего 29 кодов

- **Глубокий анализ структуры данных**

  Анализ наименования полей, их размерности, соответствия регулярным выражениям и формулам верификации.

data{crafter} Классификация

Поиск данных...

Главная | Темы | Наборы данных | Классификация | Контроль качества | Доступ | О проекте

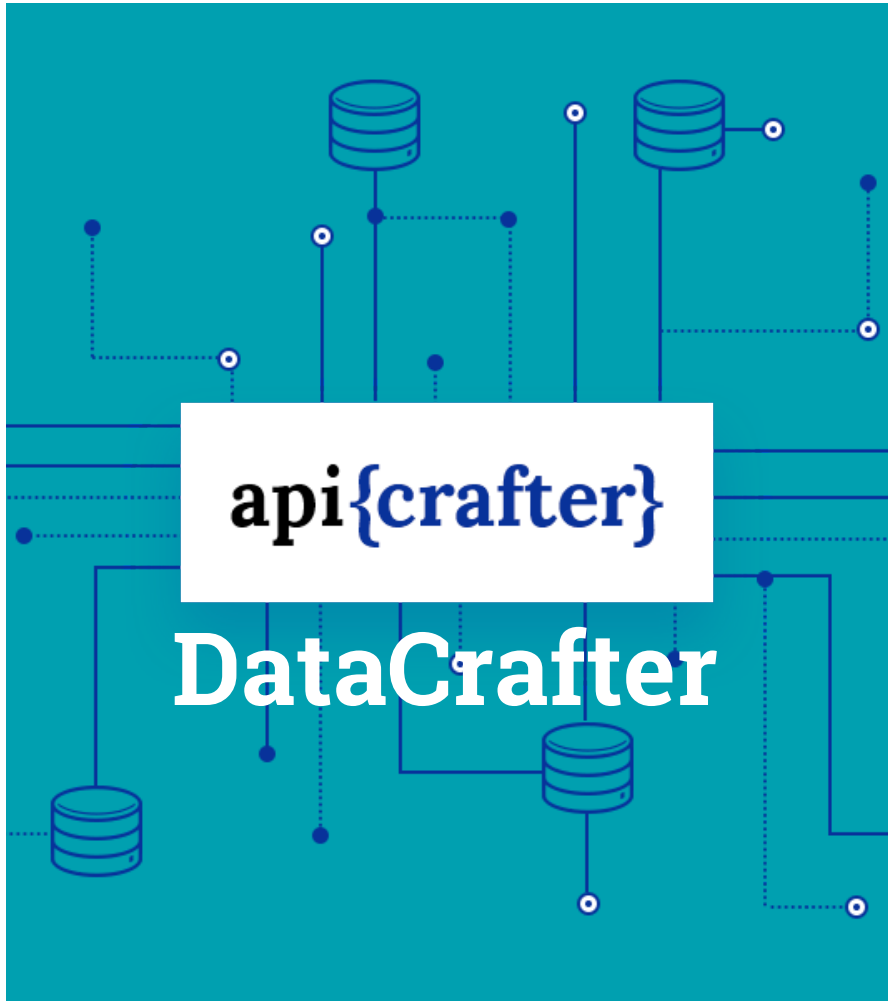
Классы данных

Код	Название	Всего полей
address	Адрес	42
bankaccount	Банковский счёт (расчётный, депозитный, корреспондентский и тд.)	1
date	Дата	3
datetime	Дата и время	198
email	Email	29
guid	Уникальный идентификатор GUID	20
inn	Идентификационный номер налогоплательщика	97
kbk	Код бюджетной классификации	0
kpp	Код причины постановки на учёт	30
ogrn	Основной государственный регистрационный номер	9
ogrn_ogrnip	Код ОГРН или код ОГРНИП	57
ogrnip	Основной государственный регистрационный номер индивидуального предпринимателя	0
okato	Код ОКАТО	7

Технологический стек

Компонент	Используются
СУБД	MongoDB и PostgreSQL
Доступ к API	REST API (Swagger/Open API) + SDK для языка Python
Экспорт данных	Форматы BSON и JSON lines
Разработка	Python

Контакты



По всем вопросам обращайтесь:

apicrafter@apicrafter.ru

Иван Бегтин