ВТБ

ВТБ / DataExchange

Решение от команды **iPrill**

Постановка задачи / Основная функциональность

- Предоставление доступа / ограничение / аудит доступа для собственных датасетов/витрин/фичей;
- Создание новых фичей из существующих данных/фичей;
- Подготовка задания на формирование датасетов по заданным правилам: правила объединения / преобразования / сэмплирования, выбор временных диапазонов.
- *Фича (feature) – переменная, сформированная из существующих данных с использованием алгоритмов преобразования

Требования

1. В качестве основного источника информации использовать *datahub api*, так же очень важно чтобы генерация заданий на формирование нового датасета / семпла осуществлялась на основании заданных правил
2. Правила должны быть в виде json-файла, их формат можно разработать самостоятельно
3. По итогам работы приложения должен быть сформирован конфигурационный файл где будут описаны все правила

Особенности нашего проекта

Особенность нашего проекта - работа с реальными датасетами взятыми по API с datahub.io.

При этом есть возможность добавлять источники (ссылки на датасеты)

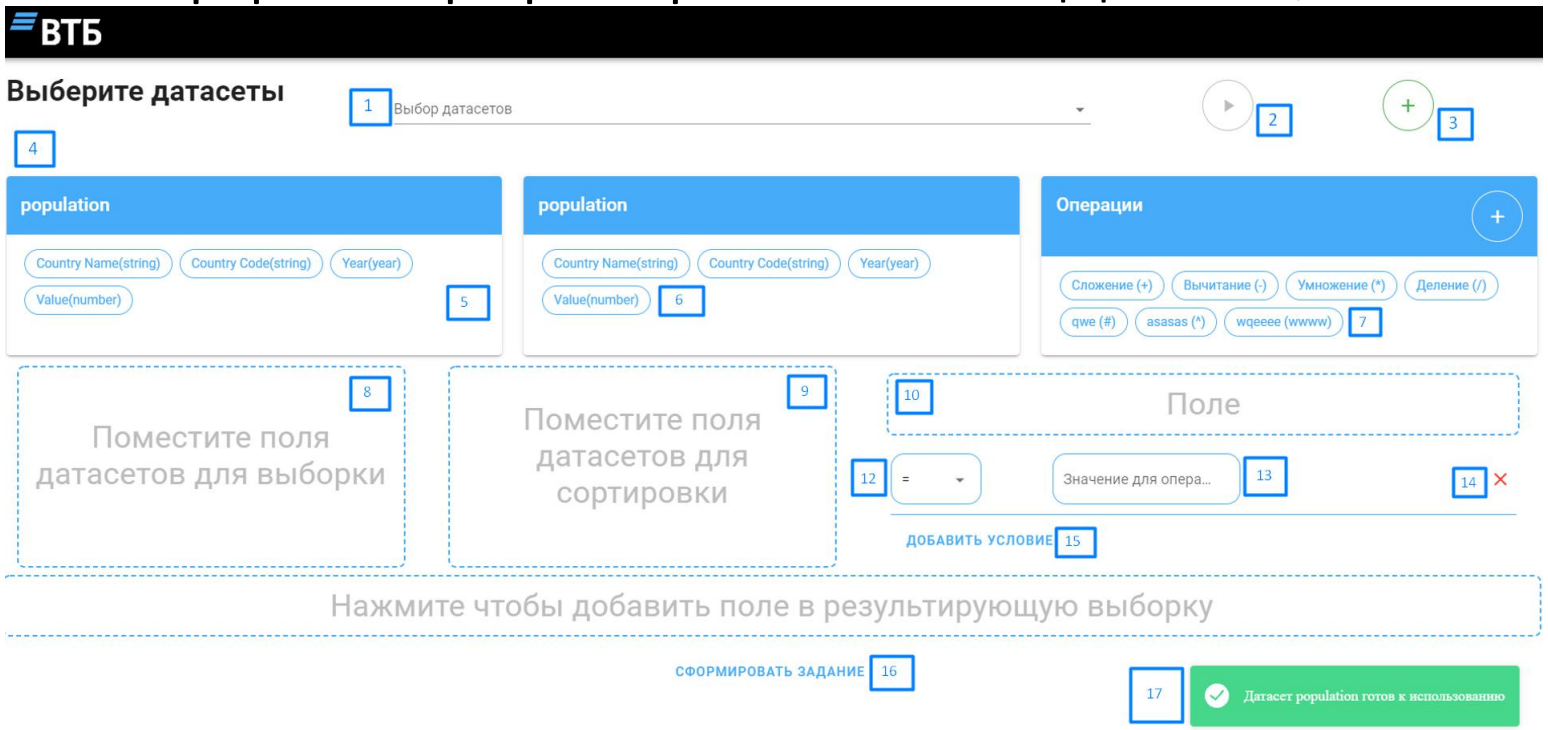
Основная идея проекта зключается в динамическом наборе полей с последующей обработкой и выгрузкой json-файла

Правила обработки набора данных задается пользователем самостоятельно через добавление результирующих полей в выборку

Правила обработки данных можно придумывать самостоятельно, затем их использовать при добавлении результирующих полей

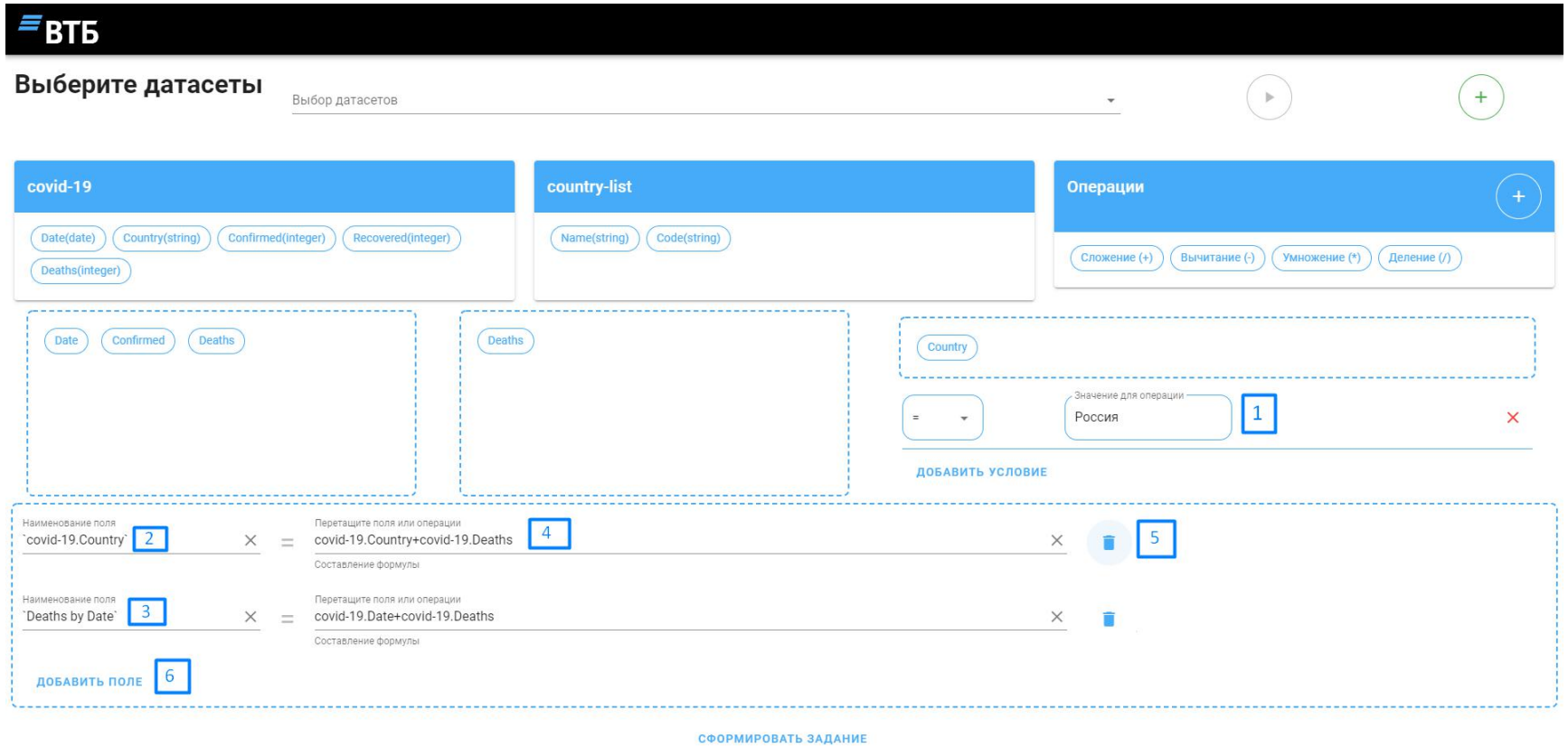
Эргономичный интерфейс формирования задания, адаптивный интерфейс

Интерфейс формирования задания / Элементы управления



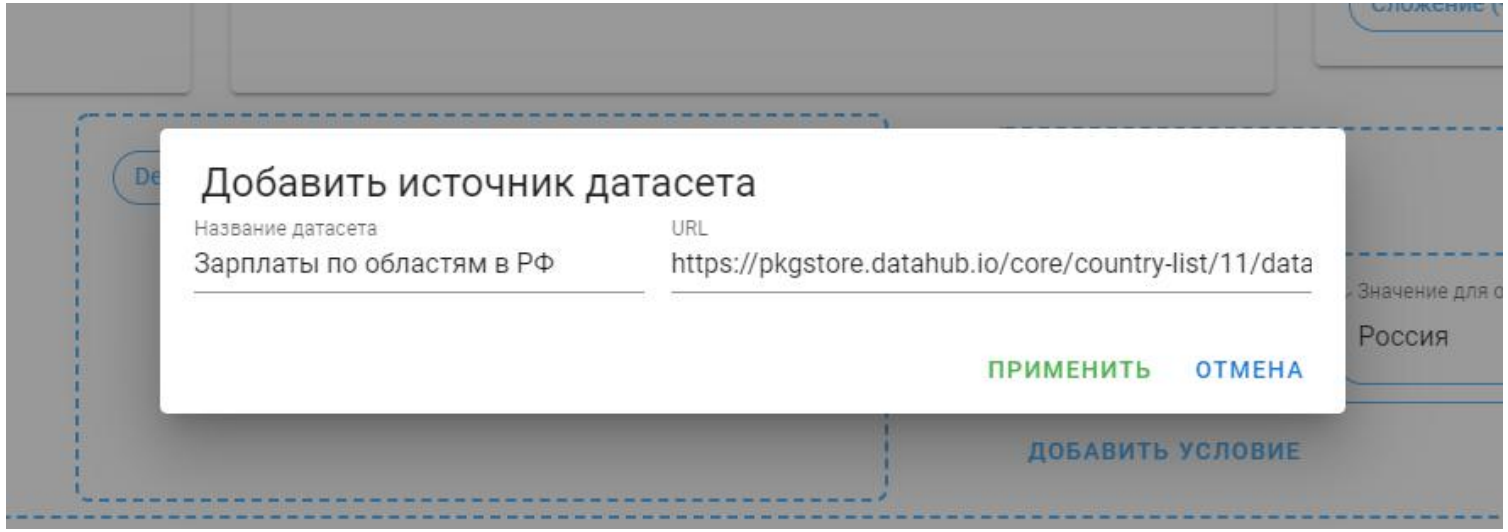
1. Выбор датасетов
2. Кнопка добавления в текущее задание выбранных датасетов
3. Кнопка добавления новых источников датасетов
4. Карточки датасетов с набором их полей
5. Поля первого датасета которые можно перетаскивать в зоны со штрихованным обрамлением
6. Поля второго датасета
7. Набор доступных операций для составления фич (*feature*)
8. Дроп-зона для перетаскивания полей датасетов которые должны использоваться в выборке
9. Дроп-зона для полей датасетов по которым должна быть произведена сортировка данных
10. Дроп-зона для перетаскивания полей датасетов по которым будет произведена фильтрация записей перед попаданием в результирующую выборку
12. Тип операции для условия фильрации по полю датасета
13. Значение для условия фильтрации по полю датасета
14. Кнопка удаления условия из задания
15. Кнопка добавления ещё одного условия в задание
16. Кнопка формирования json-файла задания для сервиса обработки данных
17. Уведомление о готовности выбранного датасета учавствовать в формировании задания

Интерфейс формирования задания / элементы управления

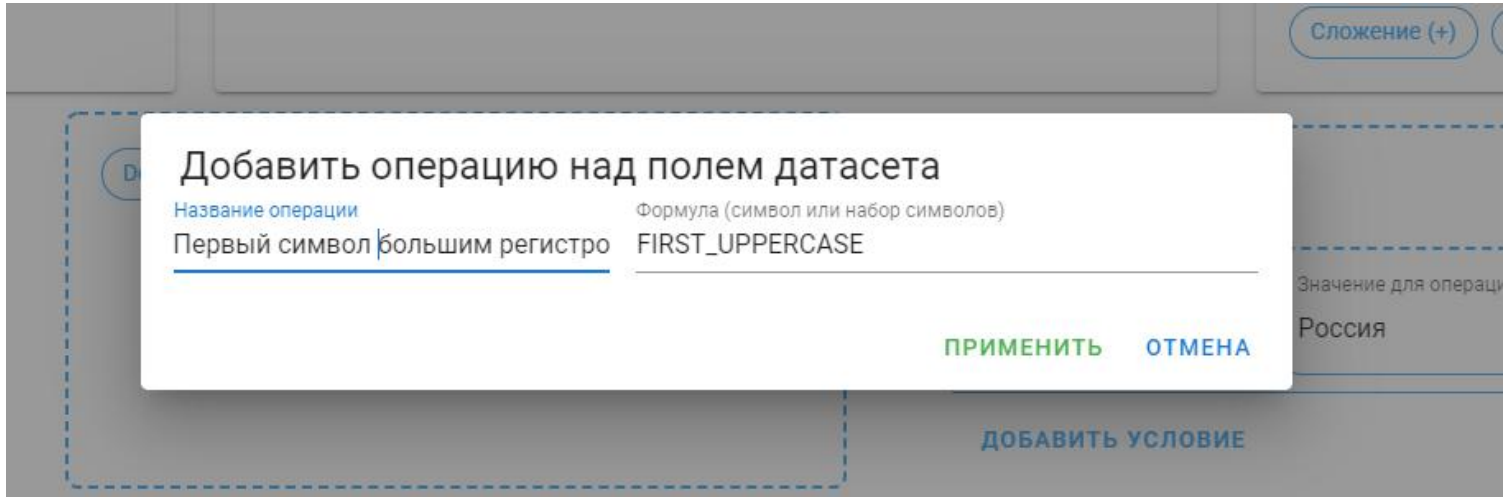


1. Значение для условия фильтрации, в данном случае фильтр по названию страны что можно интерпретировать как «Страна=Россия»
2. Наименование фичи (*feature*) - переменная, сформированная из существующих данных с использованием алгоритмов преобразования. В данном случае наименование сформировано автоматически из данных в датасете
3. Пользовательское наименование фичи
4. Формула фичи
5. Кнопка удаления фичи из текущего задания
6. Кнопка для добавления нового рабора болей фичи (добавление новой фичи)

Наполнение справочников



Окно «**Добавить источник датасета**» предназначено для добавления новых источников датасетов, с возможность его наименования. Формат json данных которые должны возвращаться для успешного добавления источника описаны на последнем слайде этой презентации в разделе «**Ресурсы**», за основу был взят формат описания датасетов сервисом datahub.io что сооветствует постановке задачи и удовлетворяет запрос на гибкость и универсальность формата описания датасетов.

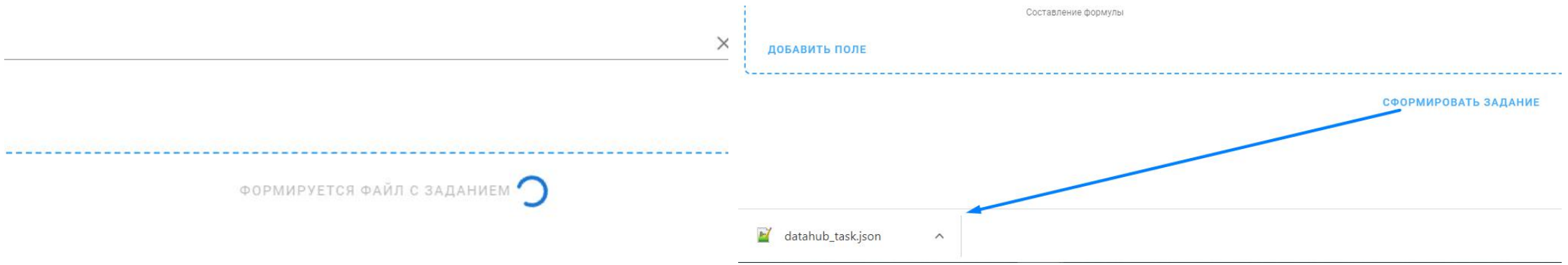


Окно «**Добавить операцию над полем датасета**» предназначено для добавления новых источников датасетов, с возможность его наименования.

Формула - это спец.слово которое будет в результирующем файле json для обозначения произвольной операции. Сервис обработки файла-задания сможет заменить это спец.слово на реальную операцию с полем датасета. Можно создавать произвольное количество пользовательских операций.

Формирование файла-задания для сервиса обработки запросов

После нажатия на кнопку «Сформировать задание» с веб-интерфейса скачивается файл на компьютер пользователя в формате json с описанием задания для сервиса по работе с датасетами.



datahub_task.json | new 33 | new 13 | new 34

```
{
    "result_fields": [
        {
            "name": "`population.Year`",
            "formula": "[population.Year]+[population.Country Code]"
        },
        {
            "name": "`population.Value`",
            "formula": "[population.Value]"
        }
    ],
    "queries": {
        "population": "select \\`population\\`.\\`Country Code\\`,\\`population\\`.\\`Year\\` from popul
    }
}
```

Описание формата JSON файла-задания

result_fields - набор полей, они же фичи (feature), из которых должен состоять результирующий датасет

result_fields[0].name - Название поля (фичи), которое содержит в себе до точки название датасета, после точки - название поля датасета

result_fields[0].formula - формула поля (фичи), которое содержит в себе инструкцию по вычислению полей из полей произвольных датасетов

queries - запросы в формате SQL для выборки, фильтрации и сортировке данных из датасетов, SQL в данном случае используется для простоты описания требуемой выборки данных.

queries.population - название датасета для которого сформирован запрос, и напротив него само тело запроса

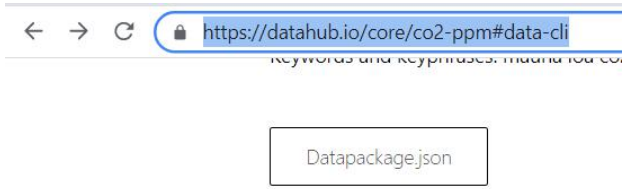
```
{
    "result_fields": [
        {
            "name": "`population.Year`",
            "formula": "[population.Year]+[population.Country Code]"
        },
        {
            "name": "`population.Value`",
            "formula": "[population.Value]"
        }
    ],
    "queries": {
        "population": "select \\`population\\`.\\`Country
Code\\`,\\`population\\`.\\`Year\\` from population ORDER BY
\\`\\`population\\`.\\`Value\\`\\`"
    }
}
```

Будущие доработки

- Валидация по типам полей при формировании результирующих полей выборки
- Добавление ценности каждой выборке и правилу (система биллинга)
- Привязка операций, датасетов и источников к пользователю
- Разграничение ролей при авторизации
- Возможность делиться датасетами и фичами
- Реализация очередей заданий через *redis* и *websocket* для асинхронной отправки заданий и получению результатов

Ресурсы

https://datahub.io/collections - страница с коллекциями наборов данных организованные по темам, на странице конкретного датасета есть возможность получить ссылку на сам файл-описание датасета (datapackage.json) и на файлы с данными по этому датасету.



https://pkgstore.datahub.io/core/covid-19/11/datapackage.json - пример json-файла с описанием датасета

https://datahub.io/docs/data-packages/tabular - документация, описание формата json-файла описывающего датасет

https://github.com/orgs/datasets/repositories - репозиторий с датасетами который можно использовать в программе, **100+ готовых датасетов**

Состав команды iPrill / Авторы программы

Поляков Андрей Сергеевич

Альметов Родион Эдуардович

Карчевский Алексей Алексеевич

Сумароков Владимир Андреевич