

Рет-проект

# Сайт «Прогноз температуры» или что я узнал за 6 месяцев

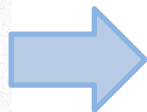
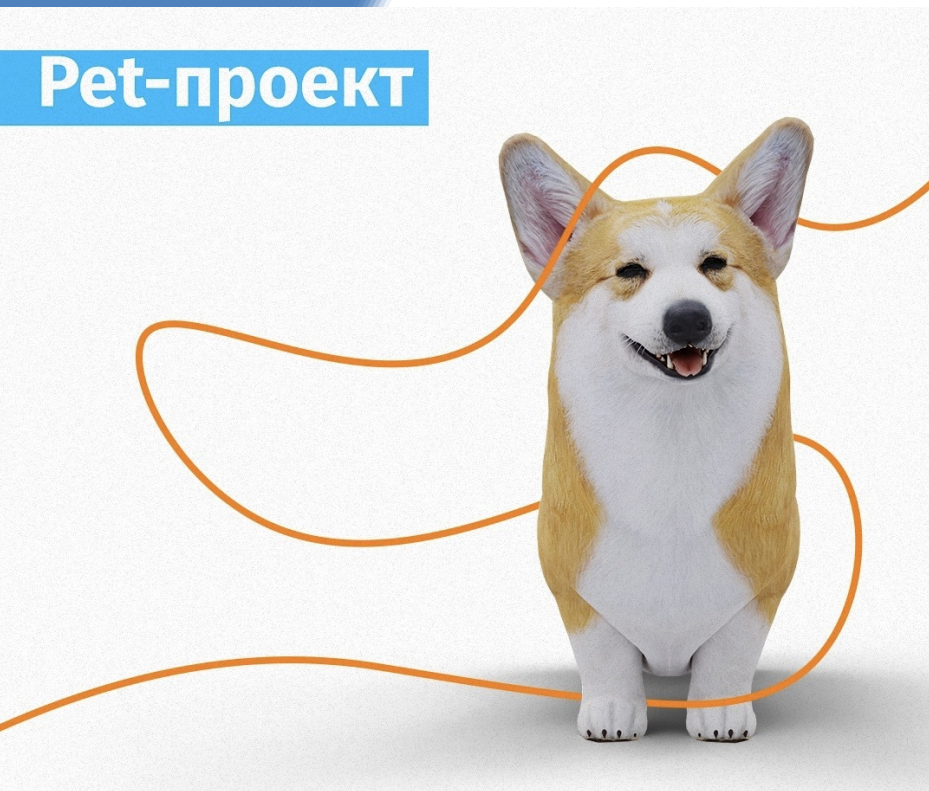
Выполнил: студент когорты 23  
Янгагин Руслан

Преподаватели: Бондарев Руслан  
Васильев Олег

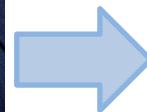
Кураторы: Мелузова Ира  
Оксана Костенюк

# Выбор задачи Pet-проекта

Pet-проект



Программист чешет репу



Что там с погодой ?  
Температура с метеодатчиков

```
Погода в Москве (ВДНХ)
2024-05-12 21:00 3.8
2024-05-12 18:00 5.0
2024-05-12 15:00 6.0
2024-05-12 12:00 6.5
```

Прогноз температуры

```
Погода в Москве (ВДНХ)
2024-05-13 00:00 2.29
2024-05-13 03:00 0.88
2024-05-13 06:00 1.81
2024-05-13 09:00 3.88
```



О, всё работает!  
Хороший мальчик!

# Задача DS

Как я бы хотел  
получать  
задачи

	DT	T
	10.05.2024 15:00	12.0
	10.05.2024 12:00	10.3
	10.05.2024 09:00	7.3
	10.05.2024 06:00	2.6
	10.05.2024 03:00	3.0
	10.05.2024 00:00	2.9

m0	m1	m2	m3	hh	T-3	T-2	T-1	T0	T+1
сезон				час	температура в истории				
0	1	0	0	3	3.0	2.6	7.3	10.3	12.0
0	1	0	0	0	2.9	3.0	2.6	7.3	10.3
0	1	0	0	21	3.4	2.9	3.0	2.6	7.3
0	1	0	0	18	5.9	3.4	2.9	3.0	2.6

1. С сайта [https://rp5.ru/Архив\\_погоды\\_в\\_Москве\\_\(ВДНХ\)](https://rp5.ru/Архив_погоды_в_Москве_(ВДНХ)) получить csv-файл погоды за год
2. Считать в DataFrame, проанализировать структуру
3. Для обучения ML-модели подготовить DataFrame:  
DT - время в формате DD.MM.YYYY HH24:MI  
T - измеренную температуру
4. Проверить задается ли DT регулярно через 3ч
5. Если обнаружены пропуски T, заполнить средним между предыдущей и последующей температурой
6. Из полученного DataFrame создать новый со столбцами:  
m0 - Признак (1/0) времени года «зима»  
m1 - Признак (1/0) времени года «весна»  
m2 - Признак (1/0) времени года «лето»  
m3 - Признак (1/0) времени года «осень»  
hh - Время первого замера (T-3)  
T-3 - Первый замер Температуры  
T-2 - Второй замер Температуры  
T-1 - Третий замер Температуры  
T0 - Четвертый замер Температуры  
T+1 - Пятый замер - Прогнозируемая температура
7. Подготовить еще 3 DataFrame с разным интервалом до времени измерения прогнозируемой температуры, где в вместо T+1 задать T+2, T+3, T+4
8. Обучить на 4-х DataFrame и выгрузить в файл объекты предсказания значения методом линейной регрессии с нормированными полиномиальными признаками степени 3



# Задача DE

This XML file does not appear to have any style information associated with it. The document

```
<?xml version="1.0" encoding="UTF-8" ?>
<feed xmlns="http://www.w3.org/2005/Atom" xmlns:rd="http://www.w3.org/1999/02/22-
xmlns:fh="http://purl.org/syndication/history/1.0" xmlns:fa="http://purl.org/atom
<id>http://rp5.ru</id>
<title xml:lang="ru">Погода в Москве (ВДНХ)</title>
<subtitle>Погода в Москве (ВДНХ). Погода на неделю в Москве (ВДНХ). Погода на де
<updated>2024-05-12T07:34:13+00:00</updated>
<link rel="self" href="http://rp5.ru/rss/5483/ru"/>
<author>
  <name>RP5</name>
  <email>support@rp5.ru</email>
  <uri>http://rp5.ru/docs/about/ru</uri>
</author>
<generator>http://rp5.ru</generator>
<category xml:lang="ru" term="Погода"/>
<logo>http://rp5.ru/images/ru/logo.png</logo>
<rights xml:lang="ru">Copyright © 000 «Расписание Погоды», 2004-2024</rights>
<fh:incremental>false</fh:incremental>
<fa:max-age>10800000</fa:max-age>
<entry>
  <id>http://rp5.ru/5483/ru#time_2024-05-12T07:34:13+00:00</id>
  <updated>2024-05-12T07:34:13+00:00</updated>
  <published>2024-05-12T07:34:13+00:00</published>
  <category xml:lang="ru" term="Погода"/>
  <title xml:lang="ru">В 10:34 на метеодатчиках</title>
  <link href="http://rp5.ru/5483/ru"/>
  <summary xml:lang="ru">
    на метеодатчиках
    <span class="was_t">было</span>
    в среднем
    <span class="t_0" style="display: inline;">+4.4</span>
    (
    <span class="t_0" style="display: inline;">+3.2</span>
    ...
    <span class="t_0" style="display: inline;">+5.7</span>
    )
    <span class="t_0" style="display: inline;">°C</span>
    . В Москве (ВДНХ) сегодня в 15:00 ожидается +7°C, без осадков, легкий ветер.
  </summary>
</entry>
</feed>
```

- С сайта <https://rp5.ru/rss/5483/ru> получить XML-документ замеров температуры
- Получить значения тегов и сохранить в **STAGE**:  
**title** - регион  
**updated** - дата записи измерений  
все значения температуры - **span** со значением параметра class = 't\_0'
- Создать детальный слой **DDS**  
Создавать словарь регионов для обработки множества регионов  
Преобразовать текстовую дату в timestampz(0)  
Отфильтровать нечисловое значение температуры, привести к числовому типу
- Создать витрину - слой **MART** содержащий :  
Справочник регионов  
Усредненное значение температуры за час в регионе  
Представление с набором признаков, на которых обучены модели  
m0 - Признак (1/0) времени года «зима»  
m1 - Признак (1/0) времени года «весна»  
m2 - Признак (1/0) времени года «лето»  
m3 - Признак (1/0) времени года «осень»  
hh - Время первого замера температуры (t\_3)  
t\_3 - Температура за 9ч до T0  
t\_2 - Температура за 6ч до T0  
t\_1 - Температура за 3ч до T0  
t0 - Последняя известная температура  
Дополнительные поля  
region - id региона  
dt\_max - час последней усредненной температуры t\_0
- Получить прогноз температуры переданными ML-моделями и сохранить результат в файл, предложить вариант его визуализации
- В Airflow создать пайплайн, выполняющий шаги передачи и трансформации данных 1-5

# Задача DevOps

1. Результаты, переданные DE упаковать в Docker-контейнер(ы)
2. Максимально автоматизировать задачу, упаковки и инсталляции для минимизации времени на этапе тестирования и внедрения решения

# Как я буду получать задачи\*

\*по версии Яндекс Практикума

Привет, Коллега! Есть время заняться новой задачей от сантиков ?

Привет! Что за задача ? Надеюсь я справлюсь...  
И, да, я в отпуске, так что время есть ))

Руководство решило осчастливить пользователей новой функцией - прогноз погоды. Аналитики обещают рост посещаемости, но это нужно еще проверять... Поэтому пока только пилотный вариант «ВДНХ». Но! Базу проектируй так, чтобы добавить другие локации без доработок

подводные камни ?

Опытный ! Скоро Сеньором станешь ))  
Погода в источнике хранится в двух вариантах:  
1. Замеры 1р / 3 часа - 0, 6, 9, 12, ... Можно выгрузить файлом за любой период. На этих данных я учу ML модели  
2. Актуальные показания с метеодатчиков. Обновляются в произвольный момент, несколько раз в час, истории нет. Тебе брать здесь, усреднять за час и выбирать 4 трехчасовые замеры к данному часу, если 13ч - формируй температуру на 4, 7, 10, 13ч.

Пока задача не выглядит фантастически сложной.  
Ничего не забыл мне сказать, чтоб потом не переделывать ?

Да, как раз собираюсь сказать: ты же Docker освоил уже ?

Ну как, освоил... С командами всё ясно и репозиторий завел, но ни одного контейнера пока не запустил

Уффф, отлично! Прямо выручил! А то и не знаю что делать.  
DevOps, который должен этим заниматься, представляешь, тоже в отпуске. Только он телефон не берет, говорят на Мальдивах...

Вот ведь ж ! Я тоже так хочу

На Мальдивы ?

Нет, хотя бы телефон не брать! ну ок, потренируюсь с контейнерами.

Не злись, зачтешь работу за Пет-проект в ЯндексПрактикуме, всё равно же будешь делать. И вот еще, нужно будет импортировать и использовать ML-объекты в Docker-контейнере. Ты сможешь, я в тебя верю )) Подробности почтой

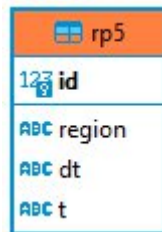
# Схема DWH

## Source

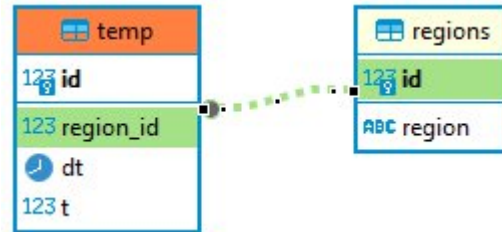
This XML file does not appear to have any style information associated with it. The document

```
<?xml version="1.0" encoding="UTF-8"?>
<feed xmlns="http://www.w3.org/2005/Atom" xmlns:rd="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:fh="http://purl.org/syndication/history/1.0" xmlns:fa="http://purl.org/atompub/facets/1.0"
<id>http://rp5.ru/</id>
<title xml:lang="ru">Погода в Москве (ВДНХ)</title>
<subtitle>Погода в Москве (ВДНХ). Погода на неделю в Москве (ВДНХ). Погода на де
<updated>2024-05-12T07:34:13+00:00</updated>
<link rel="self" href="http://rp5.ru/rss/5483/ru"/>
<author>
  <name>RP5</name>
  <email>support@rp5.ru</email>
  <uri>http://rp5.ru/docs/about/ru</uri>
</author>
<generator>http://rp5.ru/</generator>
<category xml:lang="ru" term="Погода"/>
<logo>http://rp5.ru/images/ru/logo.png</logo>
<rights xml:lang="ru">Copyright © 000 «Расписание Погоды», 2004-2024</rights>
<fh:incremental>false</fh:incremental>
<fa:max-age>10800000</fa:max-age>
<entry>
  <id>http://rp5.ru/5483/ru#time_2024-05-12T07:34:13+00:00</id>
  <updated>2024-05-12T07:34:13+00:00</updated>
  <published>2024-05-12T07:34:13+00:00</published>
  <category xml:lang="ru" term="Погода"/>
  <title xml:lang="ru">В 10:34 на метеодатчиках</title>
  <link href="http://rp5.ru/5483/ru"/>
  <summary xml:lang="ru">
    на метеодатчиках
    <span class="was_t">было</span>
    в среднем
    <span class="t_0" style="display: inline;">+4.4</span>
    (
    <span class="t_0" style="display: inline;">+3.2</span>
    ...
    <span class="t_0" style="display: inline;">+5.7</span>
    )
    <span class="t_0" style="display: inline;">°C</span>
    . В Москве (ВДНХ) сегодня в 15:00 ожидается +7°C, без осадков, легкий ветер.
  </summary>
</entry>
</feed>
```

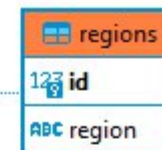
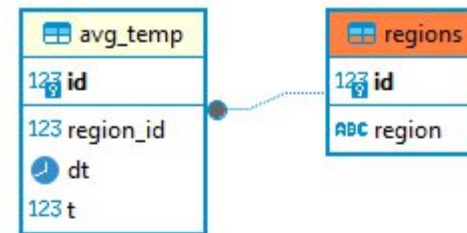
## Stage



## DDS



## Mart



data\_share\_history.txt

Погода в Москве (ВДНХ)

2024-05-13 05:00 6.3

2024-05-13 02:00 3.6

2024-05-12 23:00 3.4

2024-05-12 20:00 4.0

data share future.txt

Погода в Москве (ВДНХ)

2024-05-13 08:00 6.65

2024-05-13 11:00 9.33

2024-05-13 14:00 9.24

2024-05-13 17:00 9.59



# Общая структура решения






# Суть решения






только с собакой

# Форма решения

 **dockerhub**

Explore Repositories Organizations

ctrl+K

rus02

Search by repository name

All Content

Create repository

rus02 / weather\_yandex\_de\_ml\_server

Contains: Image • Last pushed: 4 minutes ago

Security unknown

☆ 0

↓ 4

Public

rus02 / weather\_yandex\_web\_server

Contains: Image • Last pushed: 17 minutes ago

Security unknown

☆ 0

↓ 9

Public


## Containers













[Give feedback](#)

Container CPU usage ⓘ  
2.95% / 1000% (10 cores allocated)

Container memory usage ⓘ  
2.64GB / 15.22GB

Show charts ▾

 ☒ Only show running containers

<input type="checkbox"/>	Name	Image	Status	CPU (%)	Port(s)	Last start... ▾	Actions
<input type="checkbox"/>	 <a href="#">weather_yandex_web_server</a> caf83cb89b22 	<a href="#">rus02/weather_yandex_web_server:0.1</a>	Running	0%	<a href="#">8889:80</a> 	17 minutes ago	  
<input type="checkbox"/>	 <a href="#">weather_yandex_de_ml_server</a> 03f5b7d10d2b 	<a href="#">rus02/weather_yandex_de_ml_server:0.1</a>	Running	2.95%	<a href="#">3000:3000</a>  <a href="#">Show all ports (2)</a>	12 hours ago	  

# Спасибо за внимание



# Литература

1. Курс ЯндексПрактикума "Инженер данных" 2023-2024
2. Повышение квалификации МГТУ им Н.Э. Баумана "Аналитик данных" 2022-2023
3. МЕТЕОСТАНЦИЯ с функцией прогноза погоды методом ML, Янгалин Р.Г., МГТУ им Н.Э. Баумана, 2023, BMSTU\_ML\_2022\_JRG.pptx
4. Apache Airflow и конвейеры обработки данных, Харенслак Б., де Руйтер Дж., 2021
5. Повышение квалификации "Администрирование PostgreSQL"
6. Docker Для Начинающих, Влад Мишустин, 2024, <https://www.youtube.com/watch?v=lr1rYnUubpQ>
7. Курс ЯндексПрактикума "Основы работы с Git"