

Санкт-Петербургский университет Петра Великого

# Анализ посещаемости сотрудников компании Terracor в Боливии

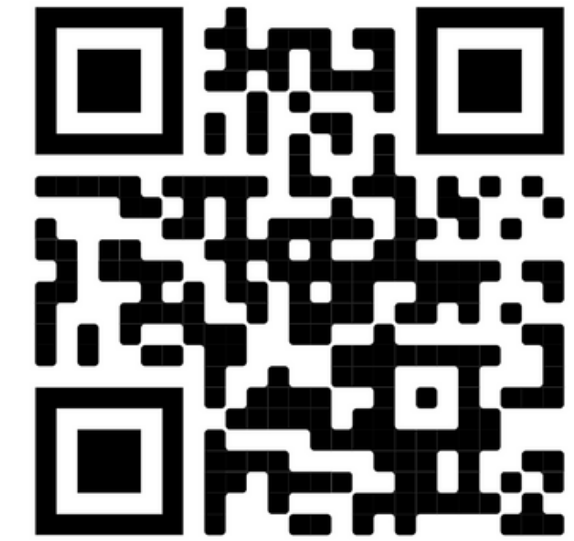
Домингес Роберто, Шушин Дмитрий,  
Бонилья Андрес, Коноплёв Николай  
5140904/40201



# 1. Анализ требований

## 1.1 Обзор предметной области

Terracor специализируется на проектировании, архитектуре, жилищном строительстве, многоквартирных домах, урбанизации, продаже недвижимости и земли.



[terracor.com.bo](https://terracor.com.bo)

### Как данный анализ может быть полезен компаниям?

#### **Улучшение управления персоналом**

Информация о принятии мер как в отношении недорабатывающих, так и перерабатывающих сотрудников

#### **Повышение экономической эффективности**

Анализ потерь от опозданий и прогулов

#### **Выявление соответствия регламентам и стандартам**

Мониторинг выполнения графика и рабочего времени для соблюдения нормативов и внутренних правил

# 1.1 Определение решаемых вопросов

## Общий анализ (компания/филиал):

- Каков общий процент посещаемости в организации?
- В каких филиалах самый высокий процент отсутствий или прогулов?
- Как посещаемость зависит от филиала за определенный период?
- В какой день недели больше всего отсутствий или опозданий?

## Индивидуальный анализ (сотрудник):

- У каких лиц самый высокий уровень накопленных опозданий?
- Сколько отсутствовал каждый человек в течение квартала?
- Как посещаемость конкретного сотрудника менялась с течением времени?
- У каких сотрудников наблюдаются повторяющиеся закономерности опозданий?

# 1.1 Определение решаемых вопросов

## Временной анализ:

- Какова общая тенденция посещаемости в течение года?
- В какие месяцы больше всего отсутствий?
- Как распределяются опоздания в часы пик (например, утренний и вечерний приход)?
- Как праздники оказывают влияние на посещаемость?

## Анализ соответствия графику работы:

- Каков уровень соблюдения графика по филиалам или группам?
- Какой процент назначенных часов был фактически отработан?
- Как среднее опоздание влияет на общую производительность?

## 1.2 Определение показателей и перспектив

### Ключевые показатели эффективности (KPI):

- Отработанные минуты.
- Минуты опозданий.
- Минуты отсутствия.
- Процент посещаемости.
- Среднее количество отработанных часов на человека.
- Коэффициент соблюдения графика.

### Перспективы:

- Человек: Индивидуальная производительность.
- Группа/секция: Внутренние сравнения.
- Филиал: Региональный анализ.
- Период времени: Сравнения между периодами (ежедневно, еженедельно, ежемесячно, ежегодно).



## 1.3 Обзор аналогичных исследований

Был проведён анализ статей по аналогичным темам, из которых хочется выделить следующие:



**Исследование и регрессионный анализ  
данных о посещаемости предприятий на  
основе технологии больших данных**

Lei Yao, Yuchuan Bian, Xiaoming Ji, Mingyu Guo -  
Пекинский институт инженерии окружающей  
среды космических аппаратов Пекин, Китай



**Анализ больших данных с  
использованием Elasticsearch и Kibana:  
корреляция рейтинга с устойчивыми  
продажами электронных товаров**

Henderi Henderi, Ranty Irawatia, Indra Indra,  
Deshinta Arrova Dewi, Tri Basuki Kurniawan



## 2. Описание источника данных

### 2.1 Формирование фактов и показателей

#### Основной факт:

Центральный факт — это посещаемость, которая регистрирует данные, полученные в результате маркировки и оценки установленных графиков.

#### Индикаторы фактов:

- Отработанные минуты.
- Минуты опозданий.
- Утвержденные и неутвержденные минуты.
- Пропущенные минуты.
- Сверхурочные.
- Процент соответствия.

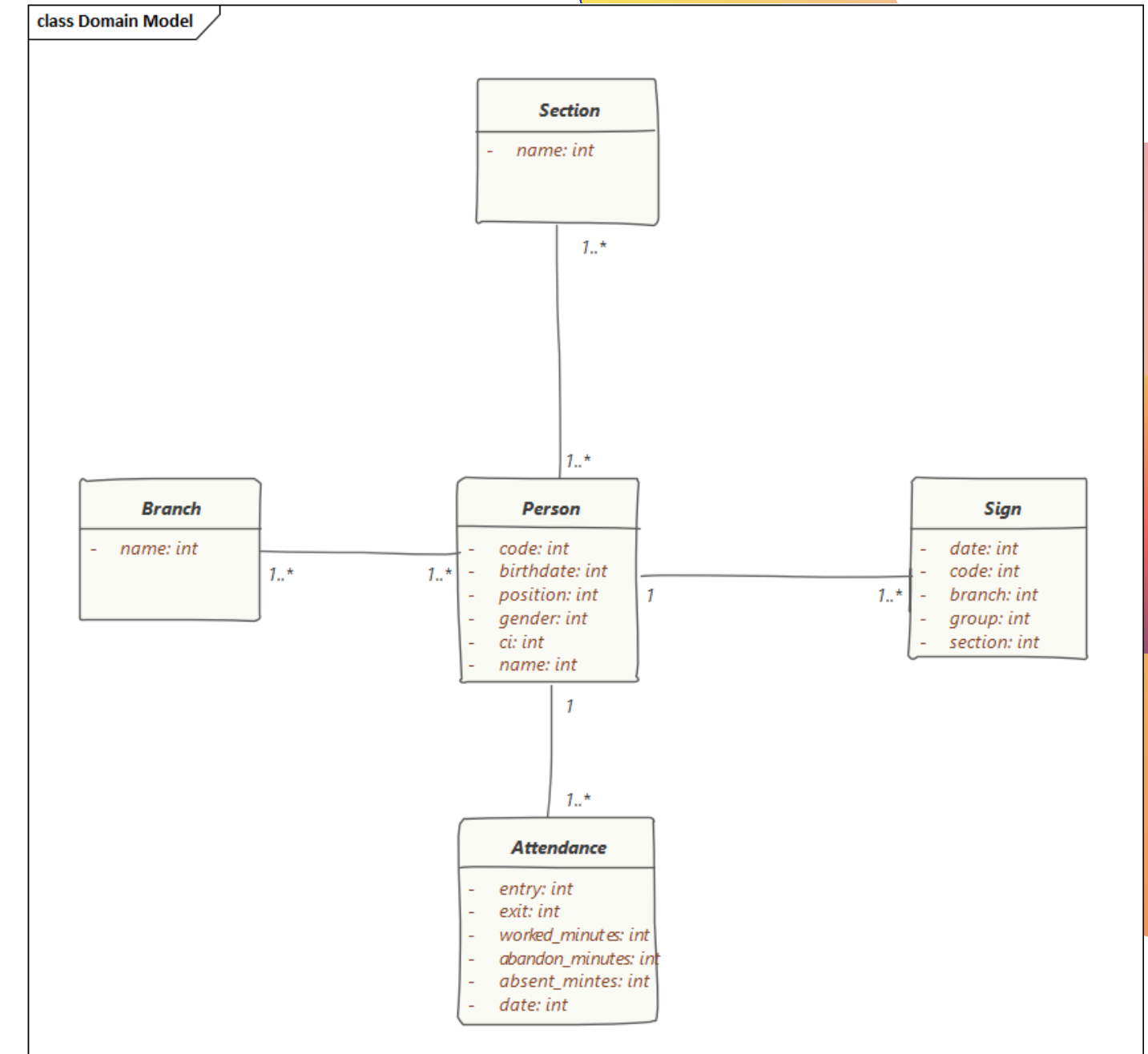
#### Связь с измерениями:

Факт связан с такими измерениями, как Человек, Филиал, Группа, Секция и Время.

## 2.2 Описание модели данных

Исходная база компании содержит 17 таблиц. Для работы был произведен маппинг данных, в результате выделены 5 таблиц, которые представляют ключевые измерения:

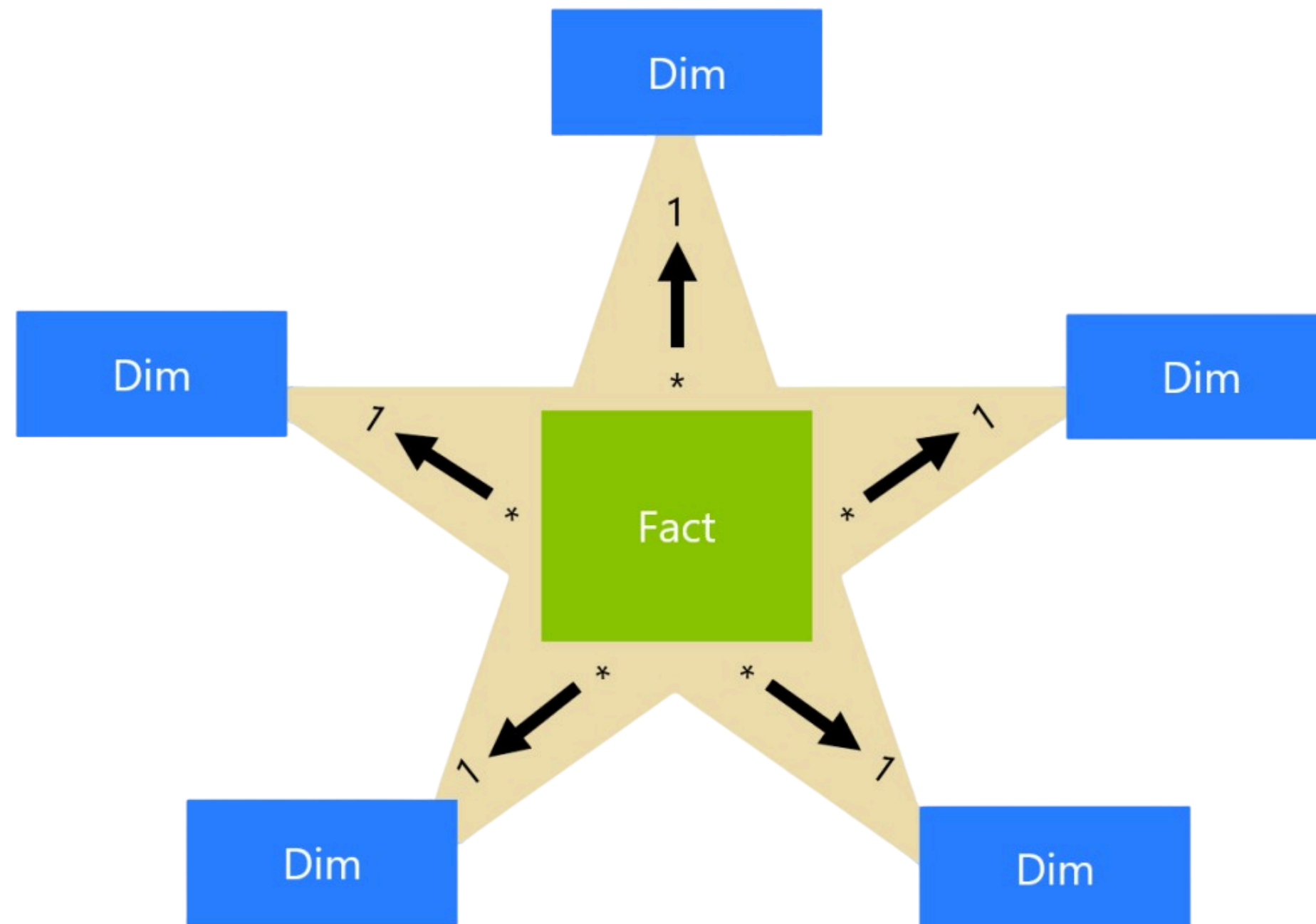
- **Person:** справочные данные, включающие общую информацию о каждом сотруднике.
- **Sign:** первичные данные, которые фиксируют фактические действия сотрудников, такие как время прихода и ухода.
- **Attendances:** данные формируются на основе отметок и содержат информацию о том, когда сотрудник присутствовал на рабочем месте.
- **Branch:** место, где производится отметка (филиал).
- **Section:** организационное подразделение сотрудников.



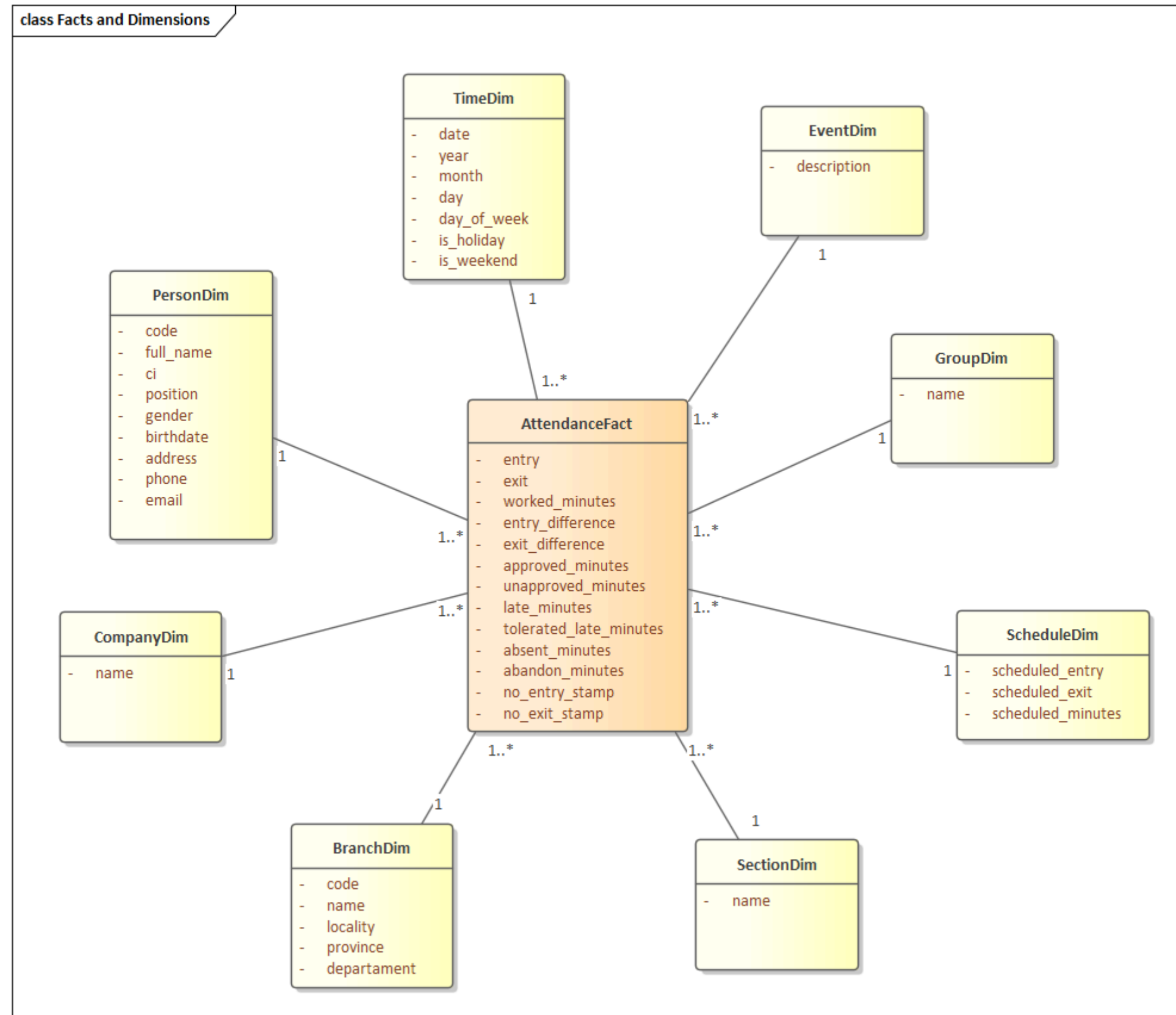


## 2.3 Концептуальная модель данных для решения

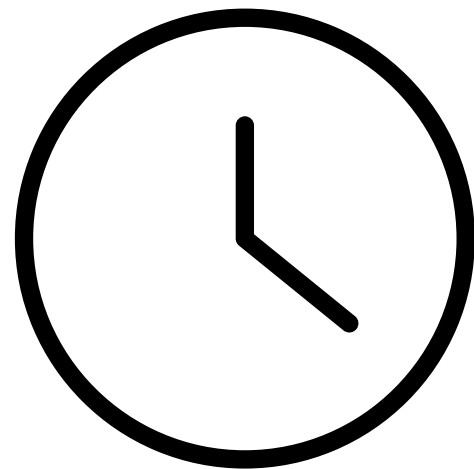
Модель данных основана на звездной топологии, но для повышения производительности запросов и данные преобразуются в денормализованный вид (единый документ с единственным основным фактом, агрегирующим ключевые измерения).



## 2.3 Концептуальная модель данных для решения



## 2.4 Ограничения в работе с источником данных



Получение данных через API занимает 1 час.



Полученные данные не структурированы, поэтому их необходимо передать в Elastic Search, чтобы объединить факты и измерения.

# 3. Архитектура решения

## 3.1 Интерфейс сбора и представления данных со стороны компании

**Интерфейс RFID:** В компании применяются физические интерфейсы, где сотрудники отмечаются с помощью RFID-меток. Используется интегрированное программное обеспечение, которое при отметке обновляет базу данных в SQL Server.

**SQL Server:** База данных компании.

**TerracorApp:** Приложение с настройками система учета.

**Terracor API:** Интерфейс для экспорта данных.



## 3.2 Используемые технологии



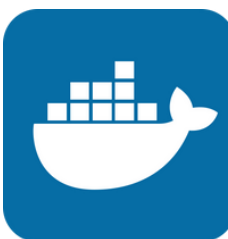
**ElasticSearch:** представляет высокоэффективную нереляционную базу данных с алгоритмом индексации



**Python:** был разработан преобразователь данных, поступающих из базы компании в Elasticsearch

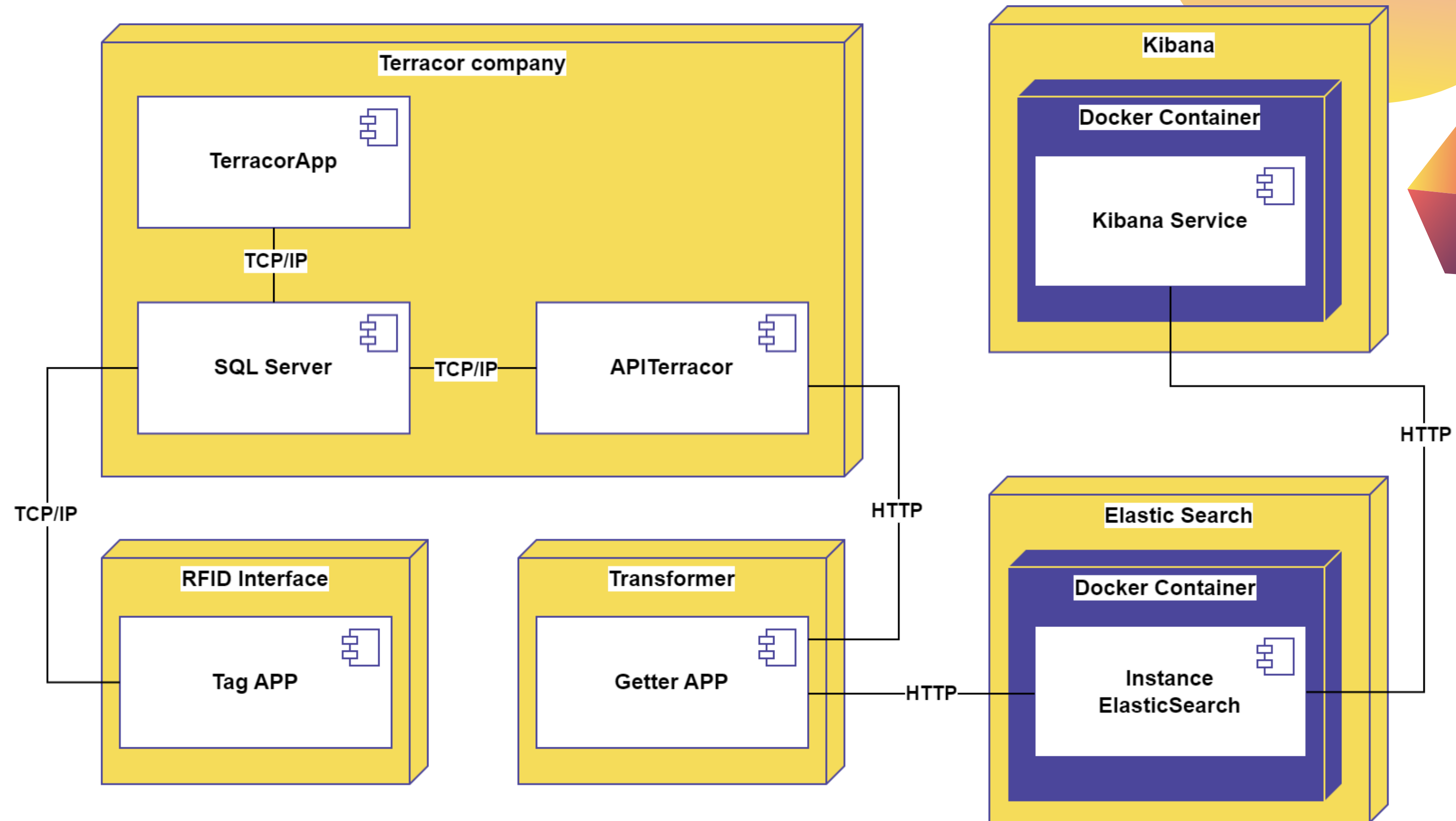


**Kibana:** предоставляет графический интерфейс, позволяющий исследовать, анализировать и визуализировать большие объемы данных в реальном времени



**Docker:** применяется для упрощения развертывания, изоляции сервисов, масштабируемости и управления версиями Elasticsearch и Kibana

### 3.3 Диаграмма развертывания системы





# 4. Работа решения

## 4.1 Особенности созданного ПО

Разработанное решение для анализа данных включает в себя несколько модулей. К ним относятся:

1. **Парсер данных.** В этом узле выполняется преобразование данных, поступающих из «Terracor API» в «Elasticsearch», с помощью программы, написанной на Python.
2. **Elastic Search.** Внутри контейнера Docker находится экземпляр сервера Elasticsearch, который представляет собой высокоэффективную нереляционную базу данных с алгоритмом индексации.
3. **Kibana.** Внутри контейнера Docker находится служба Kibana, которая подключается к базе данных Elasticsearch для выполнения соответствующего анализа данных.

## 4.1 Особенности созданного ПО

### **Terracor состоит из:**

- SQL Server: экземпляр SQL Server (база данных компании)
- Terracor App: приложение, для использования данных.
- Terracor API: интерфейс для экспорта данных.

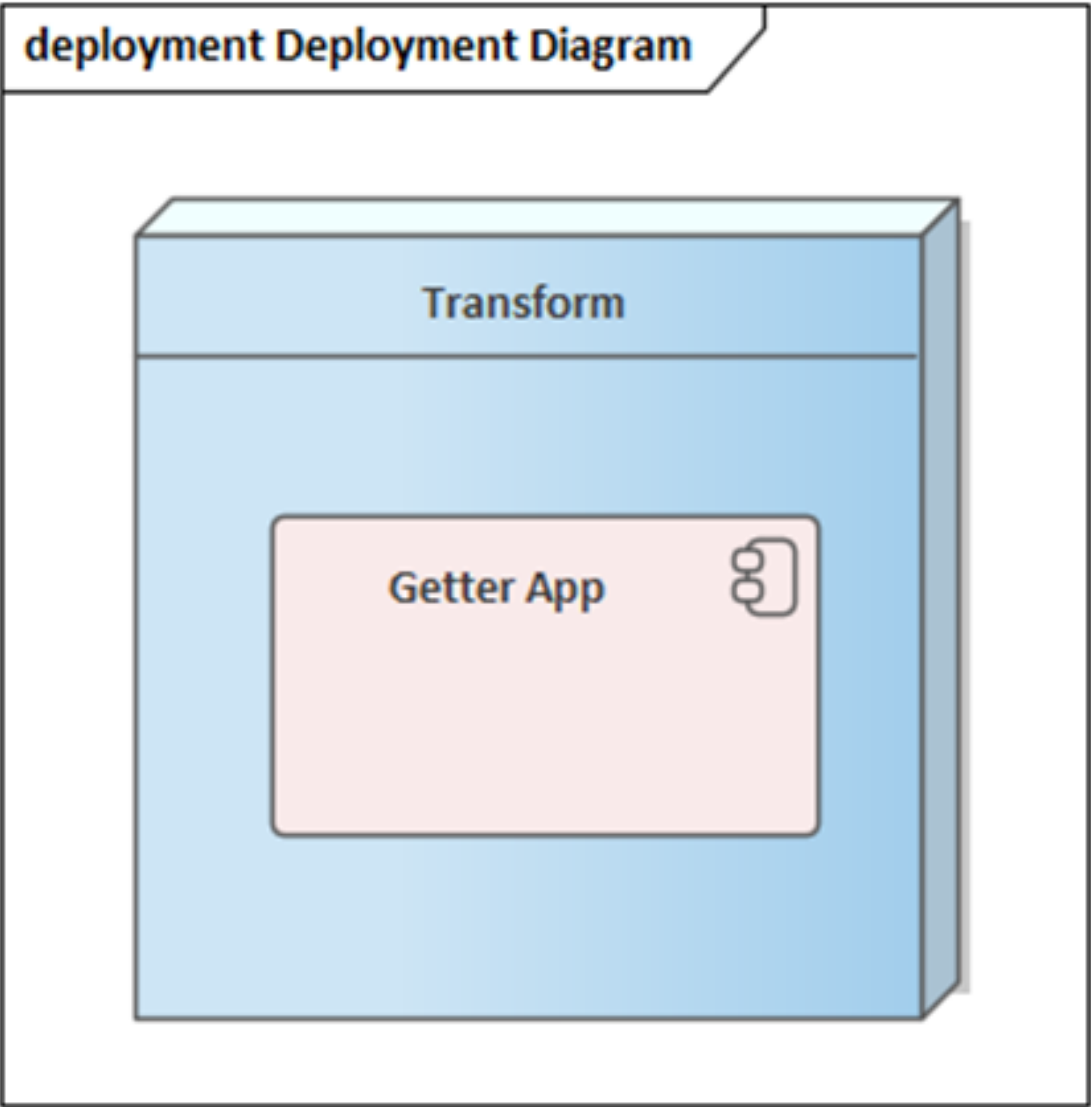
В указанной архитектуре может быть несколько компаний, поэтому наш проект хранилища данных поддерживает несколько компаний.



# 4.2 Получение данных

## Преобразование данных

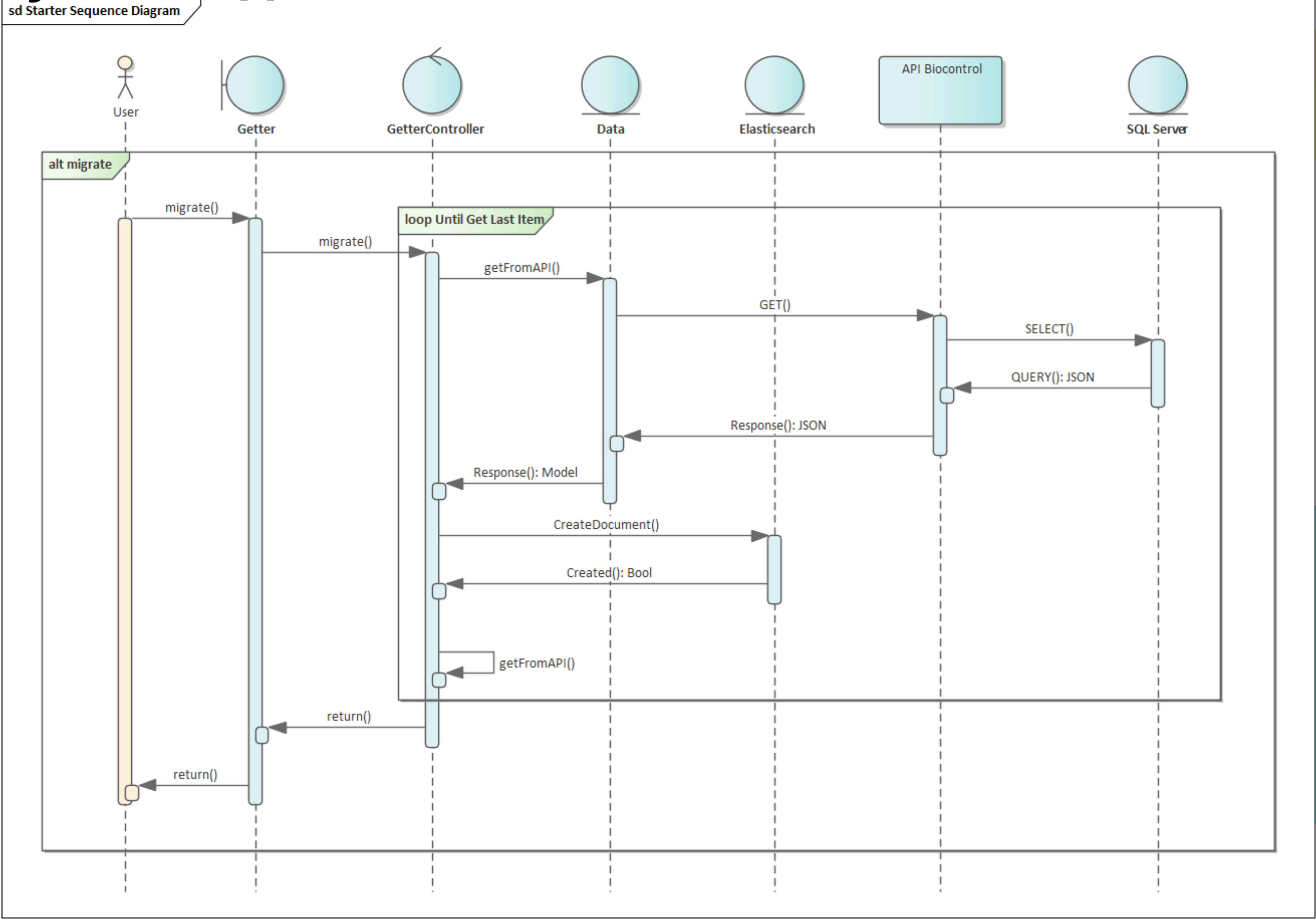
Преобразование данных выполнялось с помощью программы Python, которая использует API Terrasor, получает данные и сохраняет их во временных таблицах измерений (звездная топология), а затем передает их в денормализованную модель в Elasticsearch.



```
PROBLEMS  OUTPUT  DEBUG CONSOLE  TERMINAL  PORTS

(Attendance) Document created 1_430456
(Attendance) Document created 1_430457
(Attendance) Document created 1_430458
(Attendance) Document created 1_430459
(Attendance) Document created 1_430460
(Attendance) Document created 1_430461
(Attendance) Document created 1_430462
(Attendance) Document created 1_430463
(Attendance) Document created 1_430464
(Attendance) Document created 1_430465
(Attendance) Document created 1_430466
(Attendance) Document created 1_430467
(Attendance) Document created 1_430468
(Attendance) Document created 1_430469
(Attendance) Document created 1_430470
(Attendance) Document created 1_430471
(Attendance) Document created 1_430472
(Attendance) Document created 1_430473
(Attendance) Document created 1_430474
```

# 4.2 Получение данных



## 4.3 Горизонтальное масштабирование

### Elasticsearch

Данные распределяются между шардовыми индексами, которые можно реплицировать на несколько узлов.

### Python-преобразователь

Может быть настроен для параллельной обработки данных и отправки их в Elasticsearch

### Kibana

Взаимодействует с Elasticsearch через API, имеется возможность распределять запросы между узлами кластера Elasticsearch.

### Docker

Упрощает масштабирование сервисов, так как позволяет быстро создавать и разворачивать дополнительные контейнеры Elasticsearch и Kibana.

## 4.3 Горизонтальное масштабирование

### Принципы горизонтального масштабирования Elasticsearch:

1. Шардинг: Индексы делятся на шарды, которые распределяются по узлам. Новые узлы принимают дополнительные шарды, улучшая балансировку данных и запросов.
2. Репликация: Шарды имеют реплики на разных узлах, обеспечивая отказоустойчивость и оптимизацию нагрузки.
3. Распределение нагрузки: Запросы направляются к наиболее подходящим узлам, ускоряя обработку данных.
4. Отказоустойчивость: Размещение данных на нескольких узлах позволяет поддерживать работу даже при сбоях.

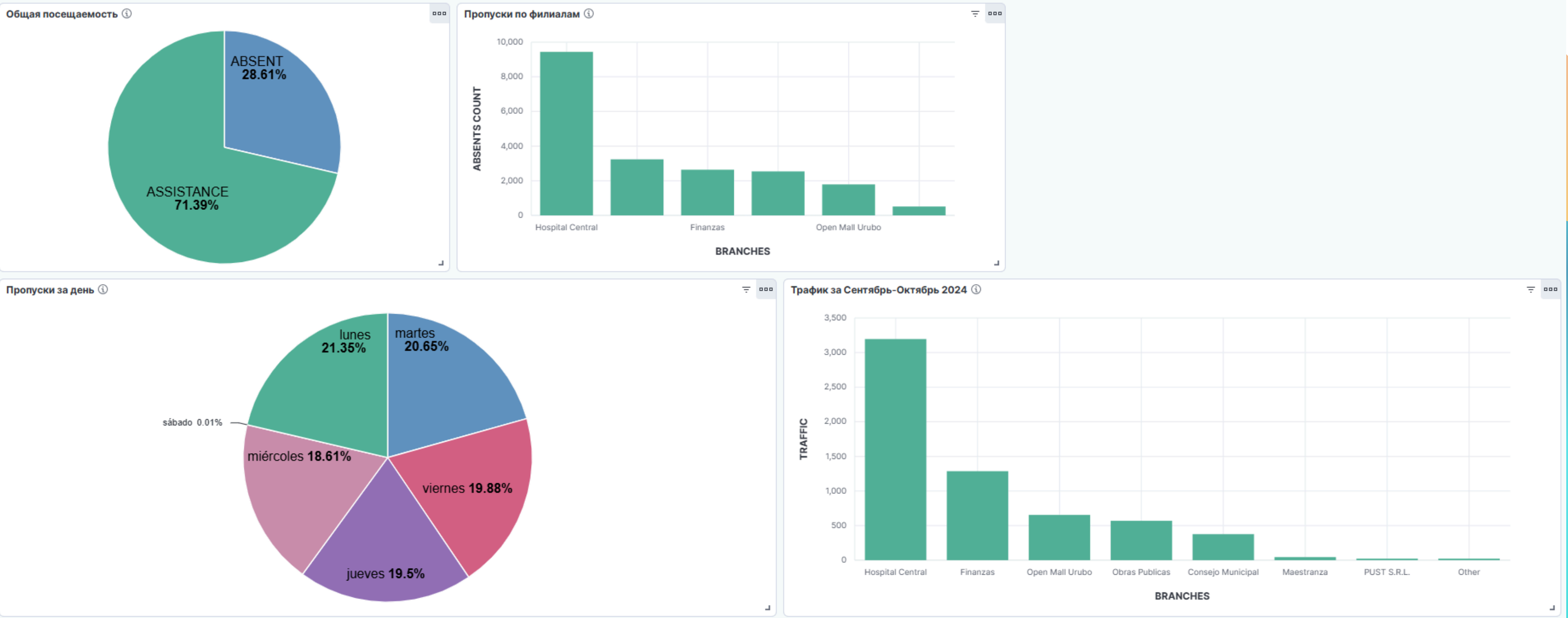


# 4.3 Временные характеристики решения

Метрика	Временные затраты	Объем
Контейнеризация Kibana и Elastic Search	10 минут	1.2 GB + 1.4 GB
Преобразование данных из базы компании	1.5 час	22 mb
Построение статистических графиков	3 часа	10 kb

# 5. Результаты анализа

## 5.1 Общий анализ (компания/филиал)

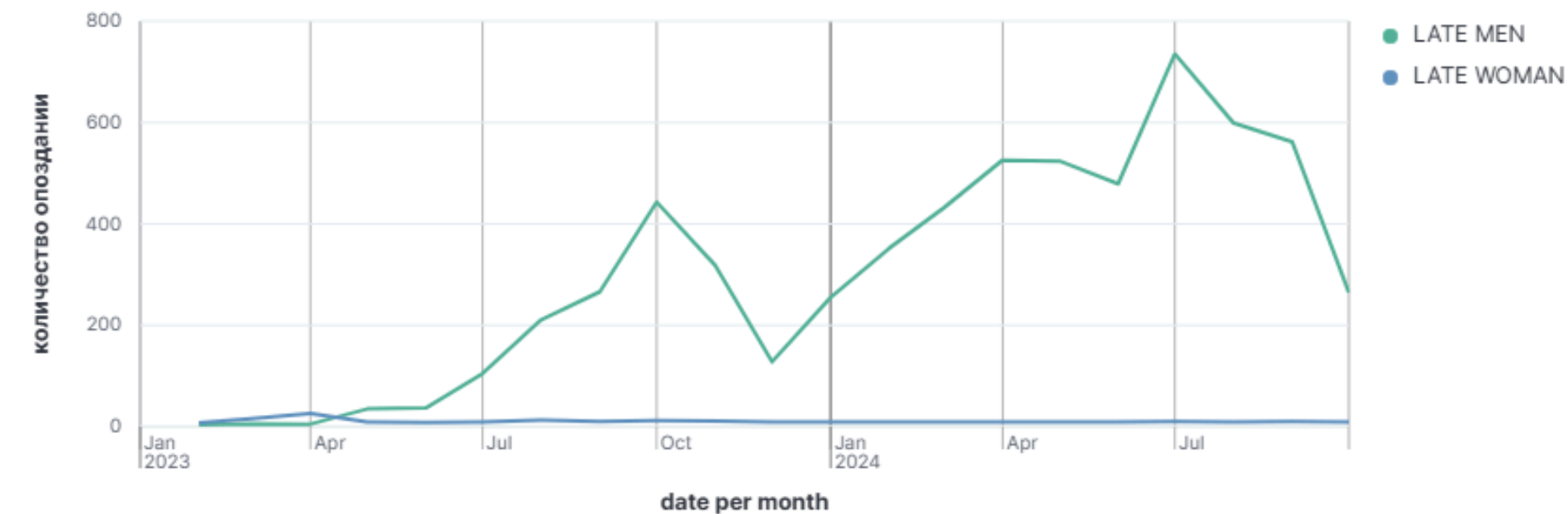


# 5.2 Индивидуальный анализ (сотрудник)

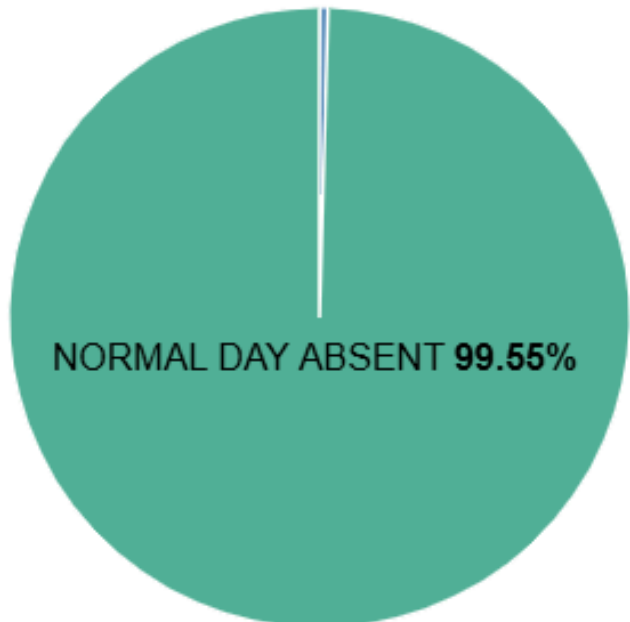


# 5.3 Временной анализ

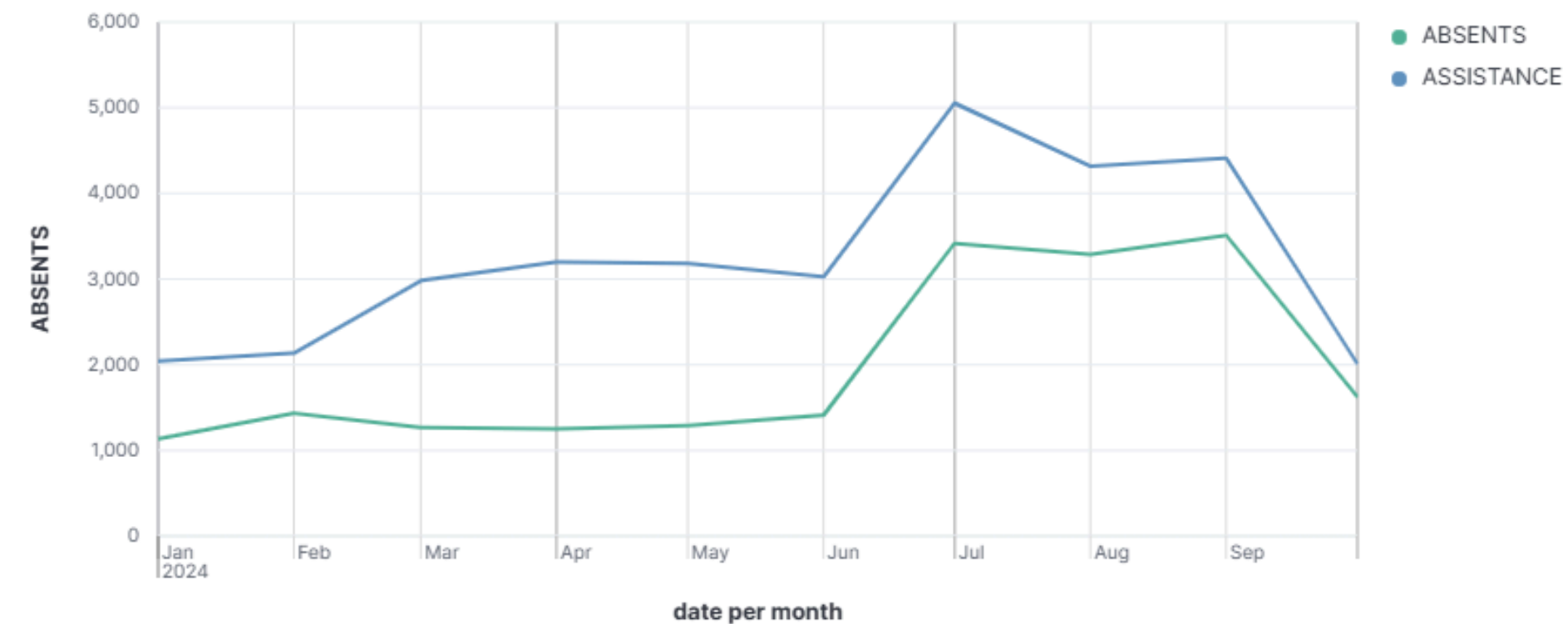
Количество опозданий по полу



Сравнение пропусков в праздничные и обычные дни



общая тенденция посещаемости 2024

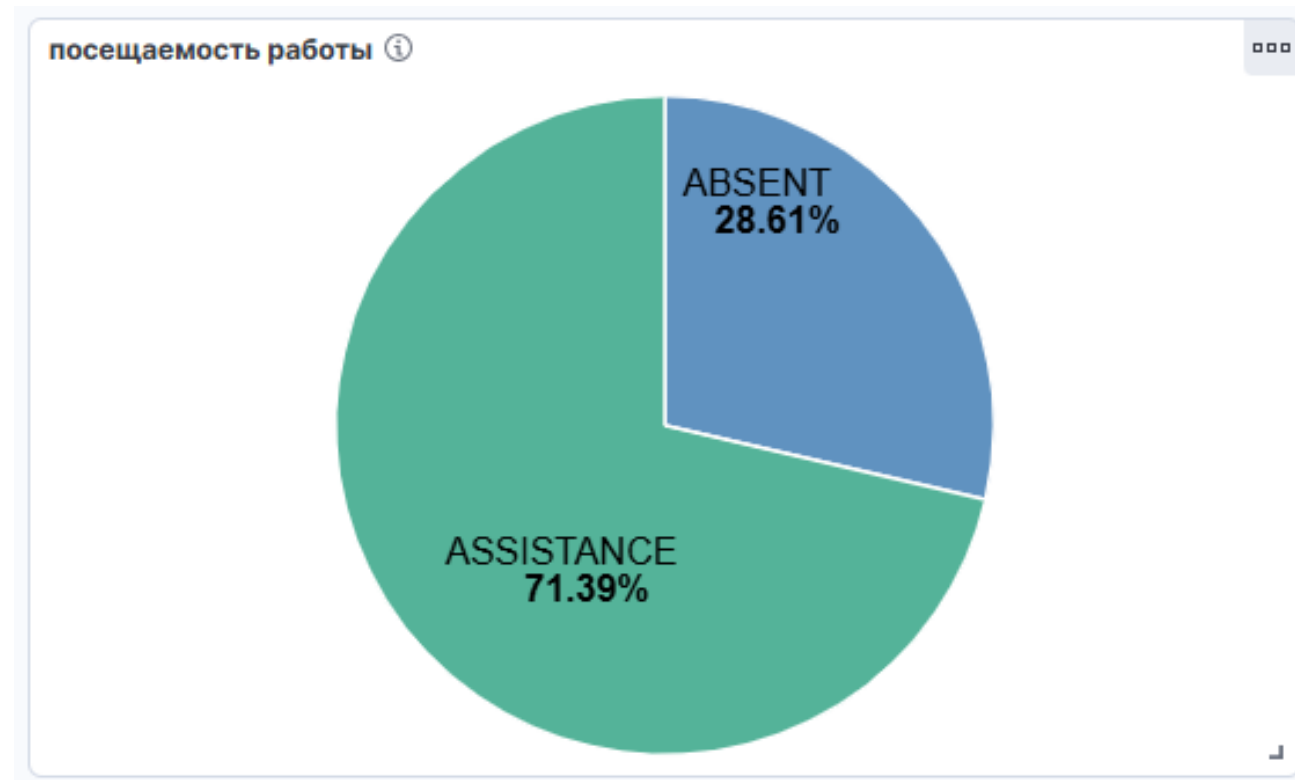


## 5.4 Анализ соответствия графику работы



## 5.5 Выводы по результатам

### Общий процент посещений



**Описание:** Круговая диаграмма или полоса, показывающая процент сотрудников, присутствовавших на своих сменах, по сравнению с отсутствующими.

**Интерпретация:** Высокий процент посещаемости отражает целеустремленность и хорошее соблюдение графика. В случае высокого процента пропусков может возникнуть необходимость проанализировать причины (пример: болезнь, отсутствие мотивации, проблемы с транспортом).

**Сравнение:** Эти данные можно сопоставить с отраслевыми показателями. Например, в исследованиях человеческих ресурсов средний уровень прогулов варьируется от 3 до 5% в зависимости от отрасли.



## 5.6 Критерии качества результатов

### 1. Percentage of Null Values:

$$\frac{\text{Null Values}}{\text{Total Records}} \times 100$$

### 2. Percentage of Out-of-Range Values:

$$\frac{\text{Out-of-Range Values}}{\text{Total Records}} \times 100$$

### 3. Mean (Average):

$$\mu = \frac{1}{n} \sum x_i$$

### 4. Standard Deviation:

$$\sigma = \sqrt{\frac{1}{n} \sum (x_i - \mu)^2}$$

### 5. Mean Absolute Error (MAE):

$$MAE = \frac{1}{n} \sum |x_i - \hat{x}_i|$$

### 6. Root Mean Squared Error (RMSE):

$$RMSE = \sqrt{\frac{1}{n} \sum (x_i - \hat{x}_i)^2}$$

### 7. Mean Absolute Percentage Error (MAPE):

$$MAPE = \frac{1}{n} \sum \left( \frac{|x_i - \hat{x}_i|}{|x_i|} \right) \times 100$$

### 8. Variance:

$$\sigma^2 = \frac{1}{n} \sum (x_i - \mu)^2$$

### 9. ANOVA (Analysis of Variance):

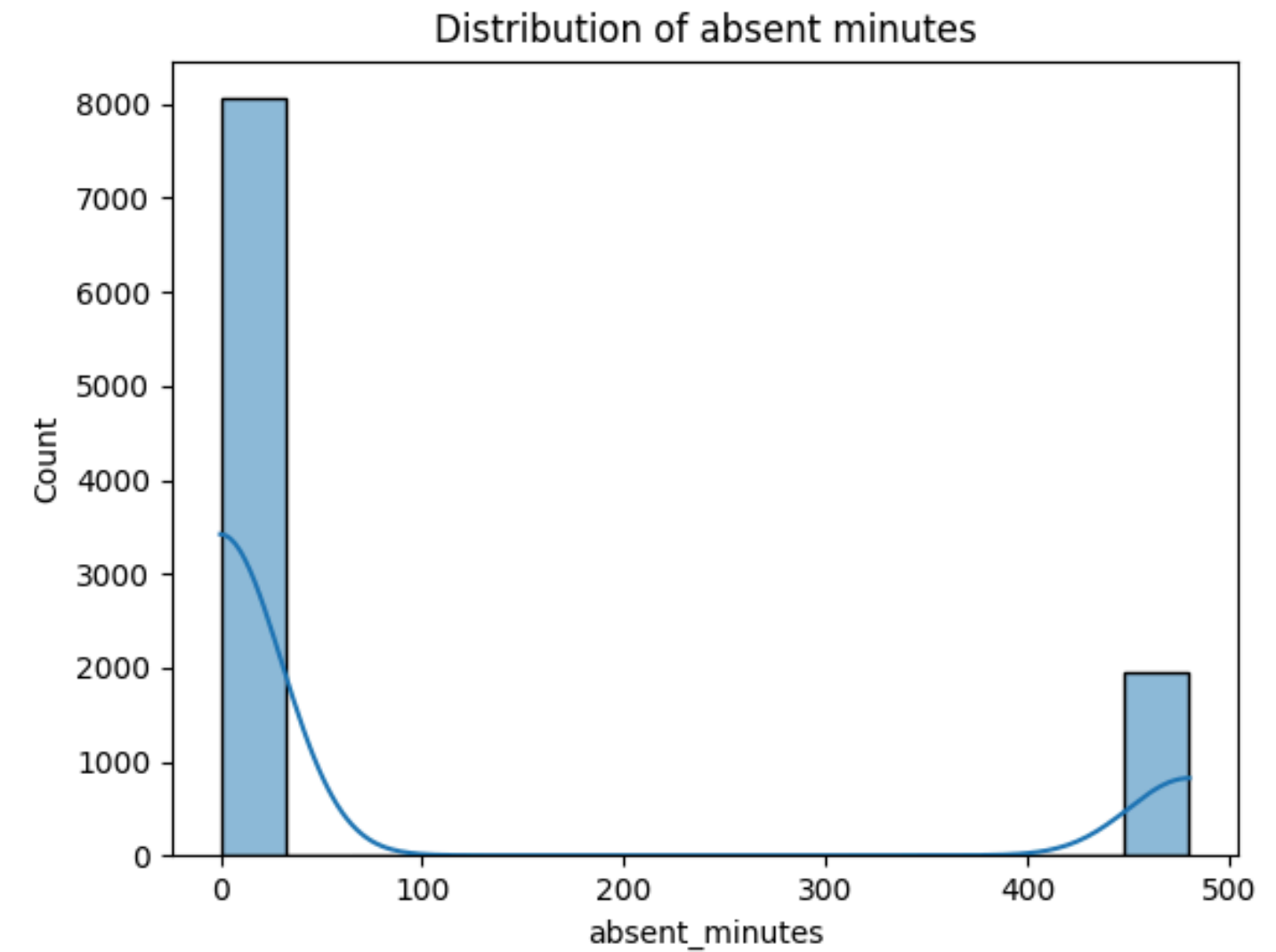
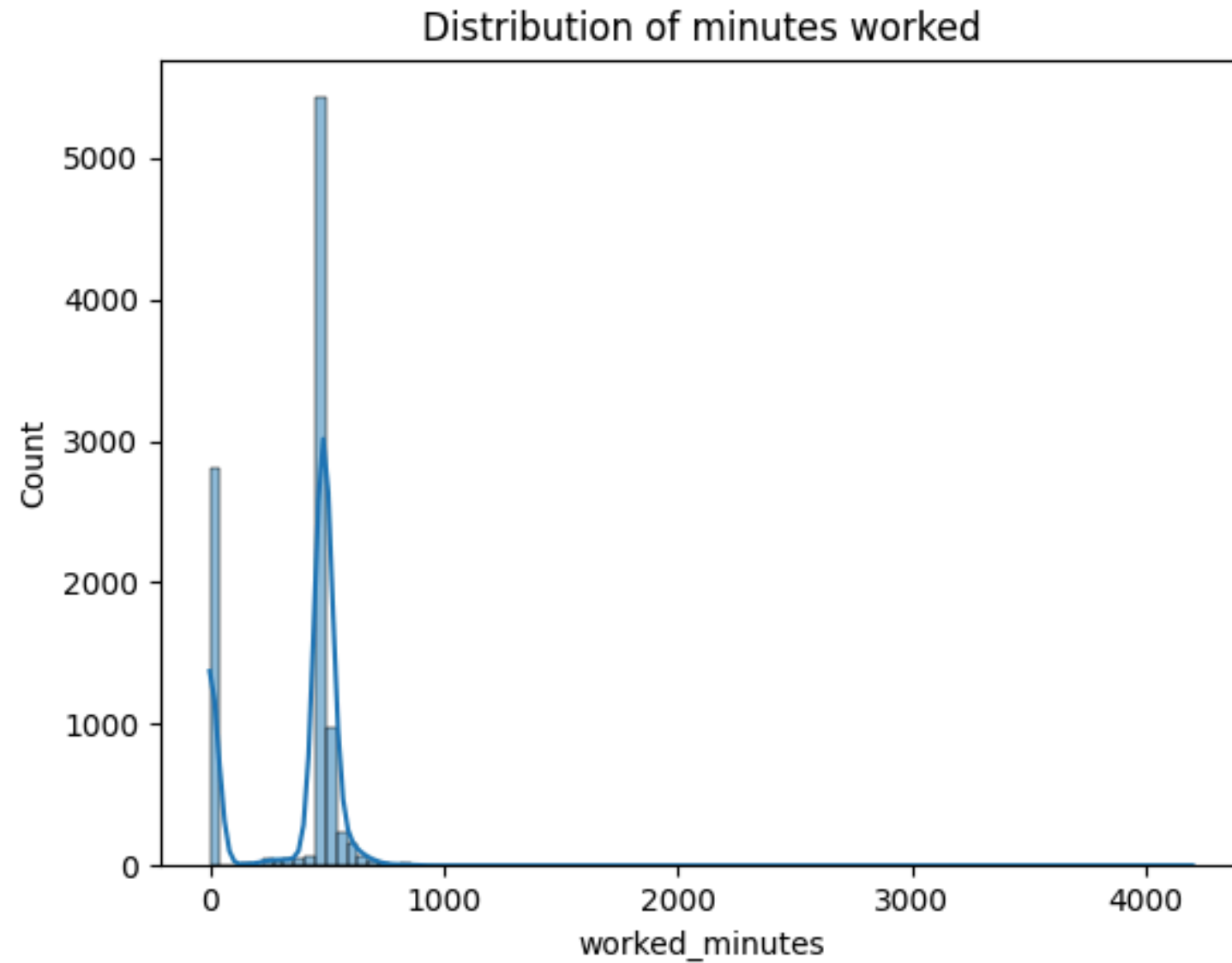
$$F = \frac{\text{Between-Group Variance}}{\text{Within-Group Variance}}$$

# 5.6 Критерии качества результатов

Missing Values:	
entry	0
exit	0
worked_minutes	0
entry_difference	0
exit_difference	0
approved_minutes	0
unapproved_minutes	0
late_minutes	0
absent_minutes	0
abandon_minutes	0
no_entry_stamp	0
no_exit_stamp	0
scheduled_entry	0
scheduled_exit	0
scheduled_minutes	0
company_name	0
date	0
year	0
month	0
day	0
day_of_week	0
is_holiday	0
is_weekend	0
event_description	0
code	0
full_name	0

ci	136
position	136
gender	0
birthdate	448
address	448
phone	448
email	448
branch_code	1088
branch_name	1088
branch_locality	1088
branch_province	1088
branch_departament	1088
branch_code	1088
branch_name	1088
branch_locality	1088
branch_province	1088
branch_departament	1088
branch_locality	1088
branch_province	1088
branch_departament	1088
section_name	9798
branch_province	1088
branch_departament	1088
section_name	9798
section_name	9798
section_locality	10000
section_locality	10000
section_province	10000
section_departament	10000

## 5.6 Критерии качества результатов



# 5.6 Критерии качества результатов

Mean Absolute Error (MAE): 152.3266  
Root Mean Squared Error (RMSE): 263.7865739570534

	worked_minutes				absent_minutes			
	mean	std	min	max	mean	std	min	max
position								
JEFE DE CONTABILIDAD	562.500000	154.568616	0	693	13.333333	80.000000	0	480
JEFE DE CONTROL, DOTACION, MOVILIDAD DE PERSON...	468.095238	139.031617	0	633	22.857143	104.744587	0	480
JEFE DE TESORERIA	335.472222	257.421443	0	678	26.666667	111.509128	0	480
AGENTE DE SEGURIDAD CIUDADANA II	0.000000	0.000000	0	0	480.000000	0.000000	480	480
ALARIFE - III	476.363636	110.195733	0	593	21.818182	102.336344	0	480
...	...	...	...	...	...	...	...	...
TECNICO DE MANTENIMIENTO DE UNIDADES EDUCATIVAS...	0.000000	0.000000	0	0	480.000000	0.000000	480	480
TECNICO DE MANTENIMIENTO VIAL	0.000000	0.000000	0	0	480.000000	0.000000	480	480
TECNICO DE RIESGO DE SEGURIDAD CIUDADANA	0.000000	0.000000	0	0	480.000000	0.000000	480	480
TECNICO OPERATIVO DE IMAGEN	410.454545	184.342520	0	584	0.000000	0.000000	0	0
VIVERISTA I	479.363636	50.214346	260	520	0.000000	0.000000	0	0

[189 rows x 8 columns]

**MAE 152,33** предполагает умеренную среднюю ошибку.

**Среднеквадратическое отклонение, равное 263,79**, указывает на то, что, хотя большинство ошибок могут быть небольшими, некоторые из них весьма велики и существенно влияют на результаты измерений.

# Дистрибутив

Исследованные данные содержат настоящие имена сотрудников, которые были обезличены для представления результатов.

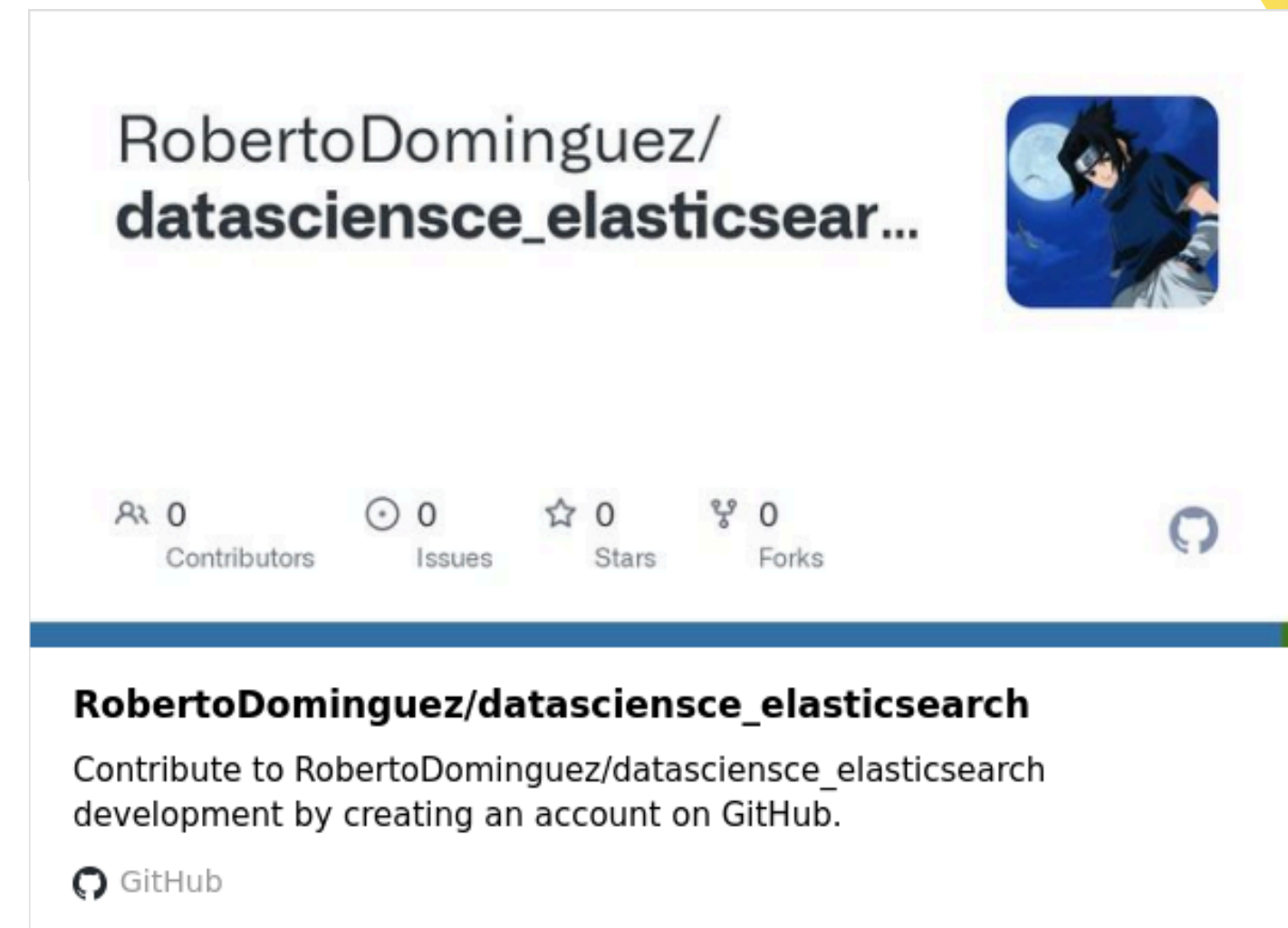
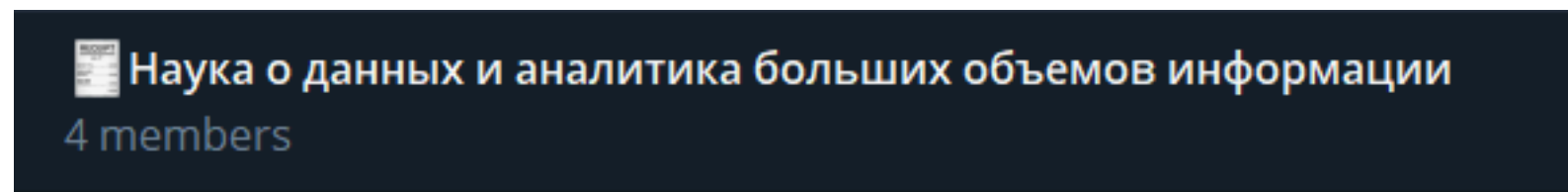


Так как эта информация конфиденциальна для компаний, она не может быть предоставлена в публичный доступ, а следовательно, решение не распространяется в виде дистрибутива.

# Приложения

технологии организации разработки программного обеспечения

- Telegram
- Github





Санкт-Петербургский университет Петра Великого

**Спасибо за внимание**

**Анализ посещаемости  
сотрудников компании  
Terracor в Боливии**

Домингес Роберто, Шушин Дмитрий,  
Бонилья Андрес, Коноплёв Николай  
5140904/40201