1. **Концепция**

В рамках реализации Задачи, состоящего из 4х этапов, предполагается сбор данных из источников, с целью использования в дальнейших расчетах, которые будут содержать данные о звонках/сообщениях в рабочей/жилой зонах, а так же уровень загрязнения в данных зонах.

1. **Обработка данных**
   1. **Источники данных.   
      Форматы данных.**

Имеются три источника данных: два datasets и карта-сетка Милана.  
Datasets:

* TELECOMMUNICATIONS\_MI   
  В данном dataset хранятся данные о звонках/сообщениях и интернет активности за два месяца(11.2013, 12.2013)  
  Расположение dataset: /user/tcld/source/TELECOMMUNICATIONS\_MI
* MI\_POLLUTION  
  В данном dataset хранятся данные о расположении датчиков загрязнения и описание этих датчиков.  
  Расположение dataset: /user/tcld/source/MI\_POLLUTION  
  Данный dataset подразделяется на две папки:
* POLLUTION\_MI -Датчики
* LEGEND -Описание датчиков

Карта-сетка Милана:

* MI\_GRID

Расположение: /user/tcld/source/MI\_GRID   
Карта – сетка представлена в формате geojson и имеет следующую структуру:  
FeatureCollection:

* Features
  + Geometry
    - 5x lat/long (4 основные коордианты + 1 замыкающая)
  + Properties
    - cellId (**Square id**)

Таблица 1. Dataset TELECOMMUNICATIONS\_MI

|  |  |  |
| --- | --- | --- |
| Поле | Тип данных | Описание |
| **Square id** | numeric | Идентификатор квадрата на карте-сетке Милана |
| **Time interval** | numeric | Временной интервал. Начало временного интервала, выраженного числом миллисекунд, прошедших с января 1970 года в UTC. Конец временного интервала можно получить, добавив к этому значению 600000 миллисекунд (10 минут) |
| **Country code** | numeric | Телефонный код страны |
| **SMS-in activity** | numeric | Активность в отношении полученных SMS-сообщений внутри **Square id**, в течение **Time interval** и отправленных из страны, определяемой кодом страны |
| **SMS-out activity** | numeric | Активность в отношении отправленных SMS-сообщений внутри **Square id**, в течение **Time interval** и отправленных из страны, определяемой кодом страны |
| **Call-in activity** | numeric | Активность с точки зрения принятых вызовов внутри **Square id**, в течение **Time interval** и совершенных из страны, определяемой кодом страны |
| **Call-out activity** | numeric | Активность с точки зрения исходящих вызовов внутри **Square id**, в течение **Time interval** и совершенных из страны, определяемой кодом страны |
| **Internet traffic activity** | numeric | Активность с точки зрения выполняемого интернет-трафика внутри **Square id**, в течение **Time interval** и по пользователям, осуществлявших соединение, определяемое кодом страны |

Таблица 2. Dataset MI\_POLLUTION/LEGEND

|  |  |  |
| --- | --- | --- |
| Поле | Тип данных | Описание |
| **Sensor ID** | numeric | Идентификатор датчика |
| **Sensor street name** | alphanumeric | Название улицы датчика. Название улицы, где расположен датчик, идентифицируемый **Sensor ID** |
| **Sensor lat** | numeric | Датчик широты. Географическая широта, указывающая положение датчика, идентифицируемого **Sensor ID** |
| **Sensor long** | numeric | Датчик долготы. Географическая долгота, указывающая положение датчика, идентифицируемого **Sensor ID** |
| **Sensor type** | alphanumeric | Тип датчика: тип датчика, идентифицируемый **Sensor ID** |
| **UOM** | alphanumeric | UOM: единица измерения значения, записанного датчиком, идентифицированным **Sensor ID** |
| **Time instant format** | alphanumeric | Формат мгновенного времени: формат, представляющий тип временной агрегации, другими словами, степень детализации данных. Это изменяется от датчика к датчику, и это может быть:  1. aggregation per day (YYYY/MM/DD)  2. aggregation per hour (YYYY/MM/DD HH24:MI) |

Таблица 3. Dataset MI\_POLLUTION/POLLUTION\_MI

|  |  |  |
| --- | --- | --- |
| Поле | Тип данных | Описание |
| **Sensor ID** | alphanumeric | Идентификатор датчика |
| **Time instant** | date | Момент времени. Момент времени измерения, выраженный в виде даты или даты / времени в следующих форматах  1. YYYY/MM/DD  2. YYYY/MM/DD HH24:MI |
| **Measurement** | numeric | Измерение. Значение интенсивности загрязнения, измеренное в момент **Time instant** датчика(**Sensor ID**). Единица измерения (UOM) значения, записанного данным датчиком, указывается в наборе данных Dataset MI\_POLLUTION/LEGEND |

**2.2 Описание этапов Задачи.**

|  |  |
| --- | --- |
| Этап | Описание |
| STG\_1 | На данном этапе необходимо учесть следующие параметры: 1. Рабочий день в Милане: 09:00 – 17:00. Обеденный перерыв: 13:00 – 15:00. 2. Нерабочим временем считать промежутки времени: 00:00-09:00 и 17:00 – 00:00 3. Square id состоит из 5 координат (5 - замыкающая)  На этапе STG\_1 нам необходимы данные из Dataset TELECOMMUNICATIONS\_MI. Необходимо суммировать всю активность пользователей (звонки, сообщения, интернет).  Определить минимальную, максимальную и среднюю активности (min activity, max activity, avg activity).   Активность от min до avg – жилая зона; активность от avg до max – рабочая зона. |
| STG\_2 | На этапе STG\_2 необходимо объединить данные этапа STG\_1 c картой-сеткой MI\_GRID для получения общей активности пользователей по **Square id.** |
| STG\_3 | На этапе STG\_3 необходимо объединить данные MI\_POLLUTION/LEGEND и MI\_POLLUTION/POLLUTION\_MI для получения информации о загрязнениях по sensor **id.** |
| STG\_4 | На этапе STG\_4 необходимо объединить данные с этапов STG\_2 и STG\_3 для получения отчета по активности и загрязнениям по **Square id.** Определить загрязненность и активность рабочей и жилой зон. |

В результате выполнения данной Задачи необходимо составить отчет со следующими данными:

- Топ 5 «загрязненных» рабочих и спальных зон.

- Топ 5 «чистых» рабочих и спальных зон.

- список не определенных по классу или по уровню загрязнения зон

В отчете сохранить данные о концентрации веществ, количестве датчиков в зоне и пороговых значениях концентрации и активности.