**Задачи DE(SQL и ETL)**

Этот набор данных содержит информацию об измерении загрязнения воздуха в Сеуле, Южная Корея. Столичное правительство Сеула предоставляет множество публичных данных, включая информацию о загрязнении воздуха, через "Площадку открытых данных".

Я сделал структурированный набор данных, собрав и скорректировав различные наборы данных, связанные с загрязнением воздуха, предоставленные столичным правительством Сеула

**Содержание**

Эти данные дают средние значения для шести загрязняющих веществ (SO2, NO2, CO, O3, PM10, PM2.5).

Данные измерялись каждый час в период с 2017 по 2019 год.

Данные были измерены для 25 районов Сеула.

Этот набор данных разделен на четыре файла.

Измерение информации измерение информации о загрязнении воздуха

после калибровки производится среднее измерение за 1 час

Measurement info.csv:

0: Нормальный, 1: Потребность в калибровке, 2: Ненормальный

4: Отключение питания, 8: Ремонт, 9: аномальные данные

**Measurement item info.csv**: Информация об элементах измерения загрязнения воздуха

**Measurement station info.csv**: Информация о приборных станциях по загрязнению воздуха

**Measurement summary.csv**: Сжатый набор данных, основанный на вышеприведенных трех данных.

* Установить СУБД Postgresql(100б)
* Загрузить csv файлы в БД ETL – инструментом(Рекомендуется Pentaho Data Integration CE).(100б)

**Написать SQL – запросы:**

* На какой станции и какого числа был зафиксирован максимальный уровень загрязненности по NO2? (100б)
* Cформировать с помощью запроса таблицу со структурой:(100б)

|  |  |  |  |
| --- | --- | --- | --- |
| **Год** | **Станция** | **Газ** | **Среднее значение** |
| 2018 | Станция 1 | CO | 0.05 |
| 2018 | Станция 1 | NO2 | 0.07 |
| ... | ... | ... | ... |

* Cформировать с помощью запроса таблицу со структурой:(150б)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **День** | **Станция** | **Газ** | **Среднее значение** | **Уровень** |
| 2018-01-15 | Станция 1 | CO | 0.05 | Green |
| 2018-02-16 | Станция 1 | NO2 | 0.07 | Yellow |
| ... | ... | ... | ... | ... |

* Реализовать все 3 запроса на компонентах ETL — инструмента. Результаты трансформации загрузить в БД, предварительно создав таблицы. Рекомендуется использовать ETL Pentaho Data Integration CE. (600б)

Итого всего(1150б)

Порог: 800б

**Задачи DS и BI**

* Сделать анализ по следующим пунктам:
* Проверка гипотез: в будние дни загрязненность воздуха ниже чем в выходные(100б)
* Прогностическое моделирование: можем ли мы предсказать концентрацию газов на основе сезонных явлении?(150б)
* Корреляционный анализ: есть ли связь между концентрацией SO2 и NO2?(100б)
* Cвободный анализ на ваше усмотрение(200б)
* Визуализировать результаты (500б). Для анализа и визуализации используйте Python

Итого всего(1050б)

Порог: 700б