**Main Learning Goals & Outcomes**

* Understand and use the relational model to structure data for efficient and effective storage and retrieval.
* Design, develop and validate a range of data models and schemas.
* Understand, evaluate and apply a range of data query and manipulation languages and frameworks.
* Constructing and reverse-engineering entity relationship models.
* Understanding and applying data normalisation.
* NoSQL [data formats and understanding the difference] to RDBMS.

**The Input Data**

The CSV files (**airquality1.csv, airquality2.csv** , **airquality3.csv**) provides air quality data ranging from 1993 to 22 October 2023 taken from 19 monitoring stations in and around Bristol.

Shown here is the first 8 lines of the file (cropped):

A screen shot of a computer

Description automatically generated

Note the following:

**There are 19 stations (monitors):**  
188 => 'AURN Bristol Centre', 51.4572041156,-2.58564914143  
203 => 'Brislington Depot', 51.4417471802,-2.55995583224  
206 => 'Rupert Street', 51.4554331987,-2.59626237324  
209 => 'IKEA M32', 51.4752847609,-2.56207998299  
213 => 'Old Market', 51.4560189999,-2.58348949026  
215 => 'Parson Street School', 51.432675707,-2.60495665673  
228 => 'Temple Meads Station', 51.4488837041,-2.58447776241  
270 => 'Wells Road', 51.4278638883,-2.56374153315  
271 => 'Trailer Portway P&R', 51.4899934596,-2.68877856929  
375 => 'Newfoundland Road Police Station', 51.4606738207,-2.58225341824  
395 => "Shiner's Garage", 51.4577930324,-2.56271419977  
452 => 'AURN St Pauls', 51.4628294172,-2.58454081635  
447 => 'Bath Road', 51.4425372726,-2.57137536073  
459 => 'Cheltenham Road \ Station Road', 51.4689385901,-2.5927241667  
463 => 'Fishponds Road', 51.4780449714,-2.53523027459  
481 => 'CREATE Centre Roof', 51.447213417,-2.62247405516  
500 => 'Temple Way', 51.4579497132 ,-2.5839890903  
501 => 'Colston Avenue', 51.4552693827,-2.59664882855  
672 => 'Marlborough Street', 51.4591419717,-2.59543271836

These monitors are spread across the four City of Bristol constituencies represented by the following Members of Parliament (MP's):

* Bristol East - Kerry McCarthy (MP);
* Bristol Northwest - Darren Jones (MP);
* Bristol South - Karin Smyth (MP); &
* Bristol West - Thangam Debbonaire (MP).

Each line represents one reading from a specific detector. Detectors take one reading every hour. If you examine the data files, you can see that the first row gives headers and there are another 1603492 (1.60 million+) rows (lines) in total in the all three files. There are 19 data items (columns) per line.

The schema for data (what each field represents) is given below:

| **measure** | **desc** | **unit** |
| --- | --- | --- |
| Date Time | Date and time of measurement | datetime |
| SiteID | Site ID for the station | integer |
| NOx | Concentration of oxides of nitrogen | ㎍/m3 |
| NO2 | Concentration of nitrogen dioxide | ㎍/m3 |
| NO | Concentration of nitric oxide | ㎍/m3 |
| PM10 | Concentration of particulate matter <10 micron diameter | ㎍/m3 |
| O3 | Concentration of ozone Concentration of non - volatile particulate matter <10 micron diameter | ㎍/m3 |
| Temperature | Air temperature | °C |
| ObjectID | Object (?) | Integer |
| ObjectID2 | Object (?) | Integer |
| NVPM10 | Concentration of non - volatile particulate matter <10 micron diameter | ㎍/m3 |
| VPM10 | Concentration of volatile particulate matter <10 micron diameter | ㎎/m3 |
| NVPM2.5 | Concentration of non volatile particulate matter <2.5 micron diameter | ㎍/m3 |
| PM2.5 | Concentration of particulate matter <2.5 micron diameter | ㎍/m3 |
| VPM2.5 | Concentration of volatile particulate matter <2.5 micron diameter | ㎍/m3 |
| CO | Concentration of carbon monoxide | ㎎/m3 |
| RH | Relative Humidity | % |
| Pressure | Air Pressure | mbar |
| SO2 | Concentration of sulphur dioxide | ㎍/m3 |

This project consists of seven tasks. This is the task breakdown:

### **Task 1: Organize and model the data**

Group the detectors by constituency and design a normalised Entity Relationship (ER) model which models all the data items.

Note that this model is a "no loss" model - that is, with the required entities holding **all** the attributes from all the derived entities.

All relationships are clearly defined and enumerated.

**output file:** An ER diagram **pollution-er.png.** Additionally, a schematic image **entitiy.png** of entities and relations.

**Task 2: Forward engineer the ER model to a MySQL database:**

Using MySQL Workbench and/or PhpMyAdmin, create the required tables and fields to hold the data. All primary and foreign key attributes are defined, and all fields have the appropriate (required) data type.

**output file:** a SQL file as **pollution.sql** showing all table and attribute definitions**.**

**Task 3: Crop and cleanse the data:**

Cleanse the cropped dataset to ensure that all dates fall between 1st January 2015 and 22nd October 2023.

**output files:** A ZIP file **cropped.zip** holding the cropped and cleansed data (CSV files for all tables in the model). Additionally, a PYTHON script **cropped.py** that accomplishes the cleaning and cropping tasks along with a screen of its output **python\_Output.png**, and an image **cleaning.png** showing cleaning and cropping steps in a table.

**Task 4: Populate the MySQL database tables with the extracted/reduced datasets created in the previous task:**

USE PhpMyAdmin’s “import CSV” feature, MySQL's “LOAD DATA INFILE” statement or Python code to import the datasets into the MySQL tables implementation completed in Task 2.

**output files:** screens capture **(4 png files)** showing the tables in the model.  
Additionally, a PYTHON script **import.py** that accomplishes the above task.

**Task 5: Design, write and run SQL queries:**

Three SQL queries for tasks below:

i) Return the date/time, station name and the highest recorded value of nitrogen oxide (NOx) found in the dataset for the year 2022.

ii) Return the mean values of PM2.5 (particulate matter <2.5 micron diameter) & VPM2.5 (volatile particulate matter <2.5 micron diameter) by each station for the year 2022 for readings taken on 08:00 A.M (peak traffic intensity).

iii) Extend the previous query to show these values for all stations for all the data.

**output files:** Three SQL queries **query-a.sql**, **query-b.sql** & **query-c.sql,** and screens **(3 png files)** of outputs.

**Task 6: Model, implement and query a selected NoSQL database.**

Model a part of data to a NoSQL data model by MongoDB to implement the selected database type/product & pipe or import a small sample of the data. Three example queries were implemented in the database.

**Submission files:** A report **nosql.md** showing 3 queries and their purposes. Additionally, screens **(3 png files)** of outputs.