Coursework Submission Requirements

***You are required to design and build a data mart/data warehouse preferably using the Oracle DBMS (Database Management System), implement required tasks. If you prefer to use an alternative DBMS then this will be acceptable as long as you can complete and demonstrate all of the requirements of this coursework. If the DBMS of your choice is not available within the university, then you will be required to demonstrate on using your own laptop.***

Students are required to read through the following scenario and design and build a data warehouse that can suitably reflect on the needs of the problem.

**Description of the Scenario**

Traditional water quality monitoring involves three steps, namely water sampling periodically, testing and investigation. This technique can be expensive, human centric, time consuming, and only provides data at the point of sampling. Water samples are taken back to the lab for analysis, or expensive handheld devices used in test locations, with each parameter requiring a specific sensor. Lab testing generates accurate spot pollutant measurements, but under normal circumstances, it is not economical for various locations all over the country. In some cases, trend of data or a close to accurate estimation of the future parameter values could be enough for businesses.

For customers the ability to remotely monitor data is the perfect solution, saving both time and money, as all they are interested in is the data, and not necessary how it is collected. Although inlocation sensors have a typical accuracy of 95% compared to lab results, this is acceptable as the customer now has real time data, trend analysis, and alerts enabling them to act much sooner than before. Multiple nodes can be deployed to build up an extensive picture of water quality. However, the in-location sensors (optical or Ionic) are expensive, require maintenance, and more importantly limited range of detectable pollutants. Customers are asking for pollutant measurement where there is no commercial sensor available, for instance **water phosphates** (organic, inorganic & total), or where sensors are expensive, **Nitrate & Nitrite**. In-location sensors are not suitable for long term deployment in water, some 6 weeks life before maintenance. Ionic sensors can be confused from other heavy ions in the water and provide false readings.

To prepare the dataset, samples are taken at sampling points around England and can be from coastal or estuarine waters, rivers, lakes, ponds, canals or groundwaters. They are taken for a number of purposes including compliance assessment against discharge permits, investigation of pollution incidents or environmental monitoring.

It has been decided to use the data from the [Department for Environment Food & Rural Affairs](https://environment.data.gov.uk/water-quality/view/landing) in the first instance. The Environment Agency use an online Data Service Platform to provide the water quality dataset to public. The Water Quality Archive provides data on water quality measurements. Samples are taken at sampling points around England and can be from coastal or estuarine waters, rivers, lakes, ponds, canals or groundwaters. They are taken for a number of purposes including compliance assessment against discharge permits, investigation of pollution incidents or environmental monitoring. The archive provides data on measurements and samples dating from 2000. The data columns for the system are given below. The collected data from 2000 until 2016 can be found on the student portal or from your course coordinator for Data Warehousing COMP1848 under the name **WaterQuality\_CW.zip**. There are many concerns about the quality of data in the database.

You should design the Data Warehouse which will provide information on the following:

* The list of water sensors measured by type of it by month
* The number of sensor measurements collected by type of sensor by week
* The number of measurements made by location by month
* The average number of measurements covered for **PH** by year
* The average value of **Nitrate** measurements by locations by year

# Design Data Mart/Warehouse

You should produce a star schema for your data mart design.

## ETL

In the first instance you will need to export the data from a Microsoft Access database into Oracle. You should then create a staging area in your own area. The data should be cleansed, and any necessary transformations carried out.

## Data Cleansing

You should plan your cleansing exercise by identifying the various types of error that you will search for (e.g. missing primary keys, missing foreign keys, misspellings, remove unnecessary records/ columns, impute missing values etc.) and describe the techniques which you used to find errors and cleanse the data.

You should show how you have used SQL for both purposes.

## Building the Warehouse

You should create and populate the fact and dimension tables for your star schema.

The FACT table and the TIME table can be populated at the same time using a cursor.

Write SQL queries on the star schema to provide the required statistical information.

**Establish connection between Oracle and Python and Extract information**

You should create a mechanism to be able to establish a connection between Python and Oracle and populate required data from your star schema in Python. Establish data preparation (e.g., table selection, query results, connection string, etc.). Additional dataset can be provided upon request. The given source code must be error free.

## Deliverables

Submit a report to support your implementation which should include:

* Explain and justify every step of your DW design and implementation in the report.
* Your Star Schema Design and BUS plan.
* Documentation for your ETL processes to include all scripts.
* A data cleansing plan together with any scripts to identify and rectify errors.
* PL/SQL code listings for your system.
* Scripts for SQL for your queries.
* Your Data Warehouse BUS plan.
* Python code used for connecting to Oracle, data pre-processing, and result summary.
* A screenshot of any forms, reports, or other GUIs.
* A discussion of any problems that you encountered and how you tried to solve them.
* Submit a .zip file of your scripts used to build and query the data warehouse.

**Attend an online demo via MS Teams (on Week 12) where you will demonstrate the implementation of your data warehouse to your tutor. Members will take turn to share their screens and present different parts of their work. The connection link and time slots for this demo will be supplied by your course coordinator.**

**For bonus points you may include some of the following features:**

* An **extended** TIME table should be created and populated.
* Write the SQL to create a materialized view to show:

**The Average of your selected sensor value by every two years**

* Write a SQL script to dump the data from the water quality table into a flat file and use SQL\*Loader to populate your data warehouse water quality table.
* Write the Create table statement that will Partition the Fact table by year.
* Create a tool that will automate the cleansing exercise.

Note: The bonus work will carry the maximum of 5 marks in total (one mark of each).

* Designing star schema including the Time dimension (10%)
* ETL: export the data from a Microsoft Access database into Oracle (5%)
* Queries (10%)
* ETL: from staging area to Dimensions and Fact tables, using cursor based on the specific water sensor (10%)
* Data cleansing (15%)
* Data Warehouse BUS plan (15%)
* Implementation of Dimensions & Fact (15%)
* Python programming used for connecting to Oracle, data pre-processing, and result summary (10%)
* Writing quality and referencing (10%)

**\* If bonus questions are attempted, the group must clearly state which members participated in them and include answers in their report (maximum 5%: each question carries 1%)**

**Group members must include the full name and student number of each member on the first page of the submitted report.**

### Grading Criteria

<50% Basic system functions are not complete. Code for populating the warehouse is not working.

50% – 59% Basic system functions work. Data from one site has been cleansed and loaded into the warehouse. The report outlines the work which has been done.

60% - 69% Data from data source has been cleansed in a staging area before building the warehouse. Basic scripts have been saved for reuse and system can be queried as specified. The report shows an awareness of implementation issues.

>=70% All scripts for cleansing, transforming and loading the system have been written and organised for reuse. Extra ‘bonus’ features have been added.

The report shows critical awareness of implementation issues.

Degree of sophistication in coding demonstrated.

For tasks allocation, refer to the coursework document (page 5).