COMP1848 (2025/26)	Data Warehousing & BI	Faculty Header ID:	Contribution: 100% of course
Course Leader: Hooman Oroojeni	Coursework: Practical Development		Deadline Date: 8 Dec 2025 5 pm
Feedback and grades are normally made available within 17 working days of the coursework deadline			

Grace period (48 hours), submission cut-off: 10 Dec 2025 5pm UK time.

Plagiarism is presenting somebody else's work as your own. It includes: copying information directly from the Web or books without referencing the material; submitting joint coursework as an individual effort; copying another student's coursework; stealing coursework from another student and submitting it as your own work. Suspected plagiarism will be investigated and if found to have occurred will be dealt with according to the procedures set down by the University. Please see your student handbook for further details of what is / isn't plagiarism.

All material copied or amended from any source (e.g. internet, books) must be referenced correctly according to the reference style you are using.

Your work will be submitted for plagiarism checking. Any attempt to bypass our plagiarism detection systems will be treated as a severe Assessment Offence.

Coursework Submission Requirements

- An electronic copy of your work for this coursework must be fully uploaded on the Deadline Date of **8 Dec 2025 at 5 pm** using the link on the coursework Moodle page for COMP1848.
- For this coursework, you must submit a single PDF document. In general, any text in the document must not be an image (i.e. must not be scanned) and would normally be generated from other documents (e.g. MS Office using "Save As .. PDF"). An exception to this is handwritten mathematical notation, but when scanning do ensure the file size is not excessive.
- For this coursework you must also upload a single ZIP file containing supporting evidence (all the scripts).
- For this coursework you must also upload a video recording (maximum 5 minutes) demonstrating the work you have done.
- There are limits on the file size (see the relevant course Moodle page).
- Make sure that any files you upload are virus-free and not protected by a password or corrupted otherwise they will be treated as null submissions.
- Your work will not be printed in colour. Please ensure that any pages with colour are acceptable when printed in Black and White.
- You must NOT submit a paper copy of this coursework.
- All coursework must be submitted as above. Under no circumstances can they be accepted by academic staff

With effective and responsible use, artificial intelligence (AI) can be a useful tool to aid learning. However, its use must be guided by principles of academic integrity and with awareness of the risks it poses, when not used with care. If you use AI in the process of undertaking your assignment, for example to create an outline of your assignment or to summarize articles, etc., you should acknowledge this by adding a declaration at the end of your work. See **https://www.gre.ac.uk/ai-guidance/students**

The University website has details of the current Coursework Regulations, including details of penalties for late submission, procedures for Extenuating Circumstances, and penalties for Assessment Offences. See http://www2.gre.ac.uk/current-students/regs

Detailed Specification

This is a group work. You must work in a group of four. Access to Virtual desktop is provided and the team can connect remotely to the Oracle server. User accounts to access Oracle service is created for all the students and the password is the same as your username. The team must arrange to have Regular meetings via MS Teams/face-to-face and work together on the coursework.
Each group/team has a clear separation of roles among members of the team.

Different roles will be required to perform different tasks. See section 'Tasks allocations between group members' on page 5 for more details

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

InkWave Publishing Ltd. is an international publisher that manages a portfolio of books, e-books, academic journals and digital magazines. Its day-to-day operations involve many moving parts: authors, agents, printing vendors, bindery facilities, distribution centres (DCs) and sales channels (online retailers, bookstores, direct subscriptions). Historically, InkWave has stored each of these data streams in separate legacy systems such as simple database for orders, spreadsheets for vendor contracts, CSV files from printing houses, and manual logs from DC managers. The data are siloed, often duplicated, and come with inconsistent formats (e.g., dates written as "12-Jan-24" vs. "2024/01/12"), missing values, or erroneous entries such as negative sales figures.

Because business decisions such as author royalty calculations, print-run planning, vendor contract negotiations, and channel profitability analysis rely on accurate, timely data, the company has identified a critical need to centralise all publishing information into a single analytical platform. The platform must clean raw source files, transform them into a consistent format, and load them into a robust data warehouse that supports fast querying for business intelligence dashboards.

The goal of this coursework is to design and implement a data warehouse that brings together all publishing materials, author details, vendor information, production metrics and sales records. Your task is to extract the raw CSV files provided, cleanse them (handle missing, duplicate or out-of-range values), transform the data into a dimensional model suitable for analytical queries, and load the results into an Oracle data warehouse. The final product should enable InkWave analysts to monitor production trends, identify bottlenecks in the supply chain, forecast future print runs, and evaluate channel performance.

Dataset

You are provided with a data set that contains production, sales and operational details from several distribution centres (DCs) of the publishing house InkWave.

The data is recorded on a daily basis and includes measurements for the following key metrics:

- PrintRun – Number of units printed at the DC on that day
- BindingCost – Total cost spent to bind/finish the printed copies
- UnitsSold – Copies sold from that DC on that day
- Revenue – Gross revenue generated by those sales
- Returns – Number of returned copies
- Temperature – Ambient temperature inside the DC (°C)
- Humidity – Relative humidity inside the DC (%)
- VendorScore – Quality rating of the vendor supplying paper/ink that day (0-10)
- ProductType – Type of product produced (Hardcover, Paperback, e-Book)

Each record also includes a number of metadata fields such as station identifiers, dates, and free text notes. To view the full list of columns please explore the provided files.

The data may include errors such as:
- Missing values – Some numeric or text fields are left blank
- Out-of-range readings – e.g., negative return counts
- Duplicate records – duplicate rows due to sync glitches
- Inconsistent formats – mixed numeric representations (commas, periods)

Your task is to design a data warehouse that cleans, consolidates and structures this information for business intelligence analysis.

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

- **Top 5 Editions by Gross Margin per Region (last quarter):** For each region in the first quarter of 2024, list which editions delivered the highest gross margin.
- **Vendor Cost Trend with Rolling Average (12-month window):** List the month-by-month binding cost per vendor and the 12-month rolling average to spot rising or falling costs.
- **Channel Mix by Edition (percentage of total units sold):** For each book/issue, list the share of sales that came from each channel.
- **Author Performance: Total Revenue vs. Average Discount:** List authors whose books bring in the most revenue(>50000) and how much discount is typically applied on their sales.
- **Detect "hot" distribution centres (high return rates & low profit):** List the "hot" DCs that have high return rates (> 5 %) AND low profit margins (< 15 %).

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 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.). 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:

- Implement a cursor summarizes pollution levels every two years.
- Produce visuals and graphs in Python that represents the sales data (e.g., trends, heatmaps).
- Write a SQL script to dump the data from the tables into flat files and use SQL*Loader to populate your data warehouse tables.
- Create a tool that will automate the cleansing exercise.

Note: The bonus work will carry the maximum of 4 marks in total **(one mark of each**).

Tasks allocations among group members:

Shared task [marks equally distributed among group members] (25%)

- Designing star schema including the Time dimension (10%)
- ETL: export the data from a Microsoft Access database into Oracle (5%)
- Queries (10%)

Individual tasks (75%)

- ETL: from staging area to Dimensions and Fact tables, using cursor based on the specific 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).