Progress Report 1

- IBP Project: Intelligent Business Performance Analytics and Decision Support Platform For Small Business
- Student: Hugh Tran
- SID: 300394597
- Reporting Date: February 9, 2026

Project Objectives

The goal of this project is to build a Business Intelligence (BI) platform that helps users understand sales performance and customer opinions. The system uses sales data and customer review data to support analysis, forecasting, and sentiment analysis. The results will be shown through a web interface and Power BI dashboards.

Work Completed So Far

1. Dataset Research and Selection

Before setting up the development environment, time was spent searching for a suitable dataset for this project. The original project proposal aimed to forecast inventory demand and suggest an appropriate inventory level. However, it was difficult to find a public dataset that met all project requirements at the same time.

The dataset was expected to support sales analysis, time-series forecasting, and text analysis of customer reviews. In addition, inventory-level data was initially considered. In practice, inventory data is rarely available in public datasets because it is often treated as sensitive business information.

As a result, the dataset selection was narrowed to focus only on sales data and customer reviews, excluding inventory data. This adjustment allowed the project to remain feasible while still supporting meaningful analysis of sales trends and customer feedback.

Several public data sources were explored during the dataset research stage, including:

- Kaggle
- Google Dataset Search
- Harvard Dataverse
- UCI Machine Learning Repository
- Data.gov

Many datasets found from these sources contained only sales data or only customer reviews. Very few datasets provided both structured sales transactions and unstructured customer feedback in a usable format.

Due to time constraints and the need to follow the project timeline, a dataset from Kaggle was selected to continue the project development. The chosen dataset is available at: https://www.kaggle.com/datasets/pruthvirajgshitole/e-commerce-purchases-and-reviews

This dataset was selected so that the core functions of the BI platform could be developed and tested without delay. Once the platform is fully functional and stable, other datasets or web-scraped data may be used in future work.

The BI platform is designed for general business intelligence purposes rather than a specific industry. Therefore, using a public and generic dataset at this stage is considered reasonable and appropriate for this project.

The selected dataset contains two main tables:

- customer_purchase_data.csv: contains individual purchase transactions made by customers, including product and sales information. Attributes are:

Column Name	Description	Data Type
`TransactionID`	Unique identifier for each transaction	int
`CustomerID`	Unique ID of the customer	int
`CustomerName`	Name of the customer	string
`ProductID`	Unique ID of the purchased product	int
`ProductName`	Name of the product	string
`ProductCategory`	Category the product belongs to	string
`PurchaseQuantity`	Number of units purchased	int
`PurchasePrice`	Unit price of the product	float
`PurchaseDate`	Date of purchase (YYYY-MM-DD)	string
`Country`	Country where the transaction occurred	string

- customer_reviews_data.csv: contains customer review text, which can be used for sentiment analysis and text mining. Attributes are:

Column Name	Description	Data Type
`ReviewID`	Unique identifier for each review	int
`CustomerID`	ID of the customer who wrote the review	int
`ProductID`	ID of the reviewed product	int
`ReviewText`	Full text of the customer's review	string
`ReviewDate`	Date the review was submitted	string

Together, these two tables provide both structured and unstructured data required for sales analysis, forecasting, and customer sentiment analysis.

2. Environment Setup

The development environment has been set up successfully. The following tools are installed and working:

- Visual Studio Code
- SQL Server Management Studio (SSMS)
- Python with a virtual environment. The environment is used to separate project libraries and avoid conflicts with other projects.
- Python connection to SQL Server using ODBC

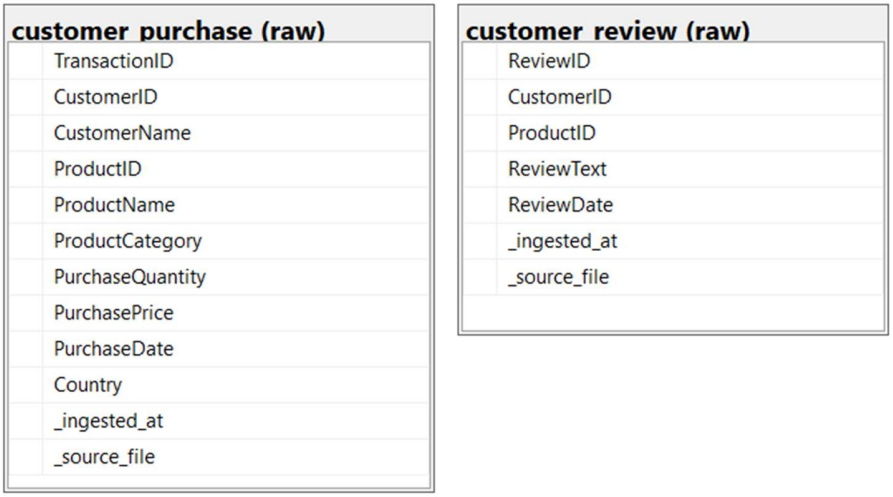
The connection between Python and SQL Server was tested and confirmed to work correctly.

```
(.venv) PS C:\Users\qhung\Videos\CSIS4495 - Applied Research  Project\W26_4495_S2_HughT\Implementation\IBP_Project> & "C:/Users/qhung/Videos/CSIS4495 - Applied Research  Project/W26_4495_S2_HughT/Implementation/IBP_Project/.venv/Scripts/python.exe" "c:/Users/qhung/Videos/CSIS4495 - Applied Research  Project/W26_4495_S2_HughT/Implementation/IBP_Project/src/test_db_connection.py"
Connected as: HTTK\qhung
Database: IBP
```

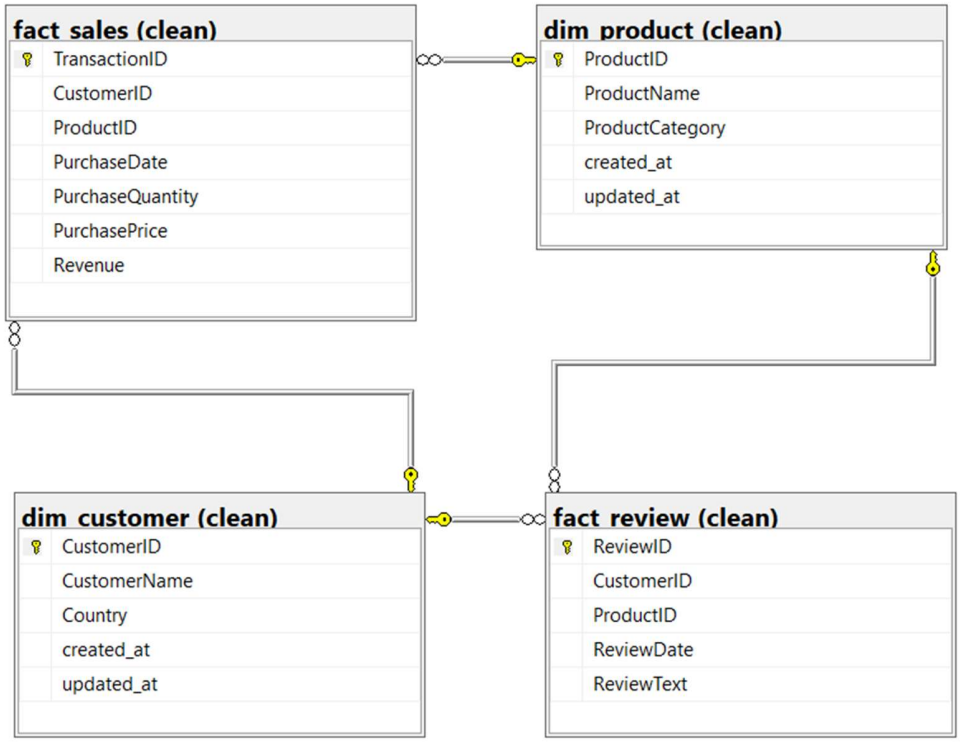
3. Data Architecture Design

A three-layer data architecture was designed and implemented:

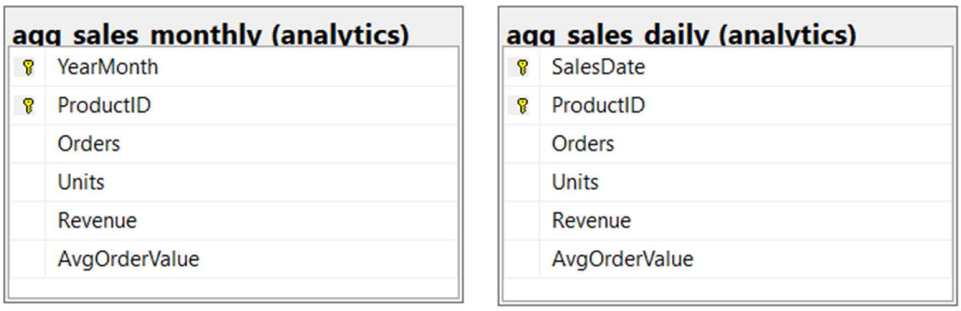
1. **Raw layer** – stores original data without changes



2. **Clean layer** – stores structured and cleaned data using tables



3. **Analytics layer** – stores aggregated data for reporting



This structure helps keep the data organized and easy to manage.

4. Raw Data Loading

Two datasets were loaded into the Raw layer:

- Sales (customer purchase) data
- Customer review data

Python scripts were used to load the CSV files into SQL Server tables:

- raw.customer_purchase
- raw.customer_review

Each table contains 1,000 records, which confirms the data was loaded successfully.

Results Messages

	TransactionID	CustomerID	CustomerName	ProductID	ProductName	ProductCategory	PurchaseQuantity	PurchasePrice	PurchaseDate	Country
1	1	887	Kenneth Martinez	240	Router	Electronics	5	689.99	2024-03-01	Barbados
2	2	560	Joseph Anderson	299	Camera	Electronics	4	79.27	2024-01-26	Northern Mariana Islands
3	3	701	Vincent Reynolds	207	Electric Kettle	Home Appliances	3	666.75	2024-05-13	British Virgin Islands
4	4	630	Christopher Morris	290	Smartwatch	Electronics	5	316.19	2023-09-21	Guatemala
5	5	631	Sarah King	281	Toaster	Home Appliances	4	700.24	2024-01-25	Falkland Islands (Malvinas)

	ReviewID	CustomerID	ProductID	ReviewText	ReviewDate	_ingested_at	_source_file
1	1	486	267	So impressed by the quality. This product truly d...	5/12/2024	2026-02-04 04:54:05.4328368	customer_reviews_data.csv
2	2	810	246	I'm very happy with the performance. It does exa...	3/7/2024	2026-02-04 04:54:05.4359123	customer_reviews_data.csv
3	3	855	291	I regret buying this. The quality is terrible and it s...	11/15/2023	2026-02-04 04:54:05.4372689	customer_reviews_data.csv
4	4	524	235	It serves its purpose, but it's not anything extraor...	5/27/2024	2026-02-04 04:54:05.4392741	customer_reviews_data.csv
5	5	238	220	Very high-quality product. I would buy it again wit...	9/10/2023	2026-02-04 04:54:05.4392741	customer_reviews_data.csv

	purchase_rows
1	1000

	review_rows
1	1000

5. Clean Data Layer

A relational database model was created using dimension and fact tables:

Dimension tables

- Customer
- Product

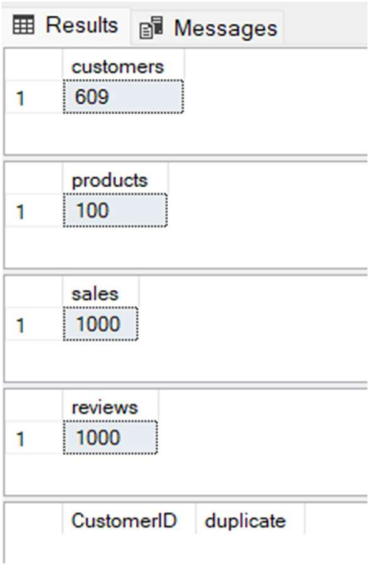
Fact tables

- Sales transactions
- Customer reviews

The tables were created in SQL Server with primary keys and foreign keys. Python was used to clean the data and insert it into these tables. This included:

- Converting dates and numbers
- Removing duplicate records
- Ensuring correct relationships between tables

After this step, the clean data model was ready for analysis.



6. Analytics Layer

The analytics layer was created to support reporting and future forecasting.

The following tables were created:

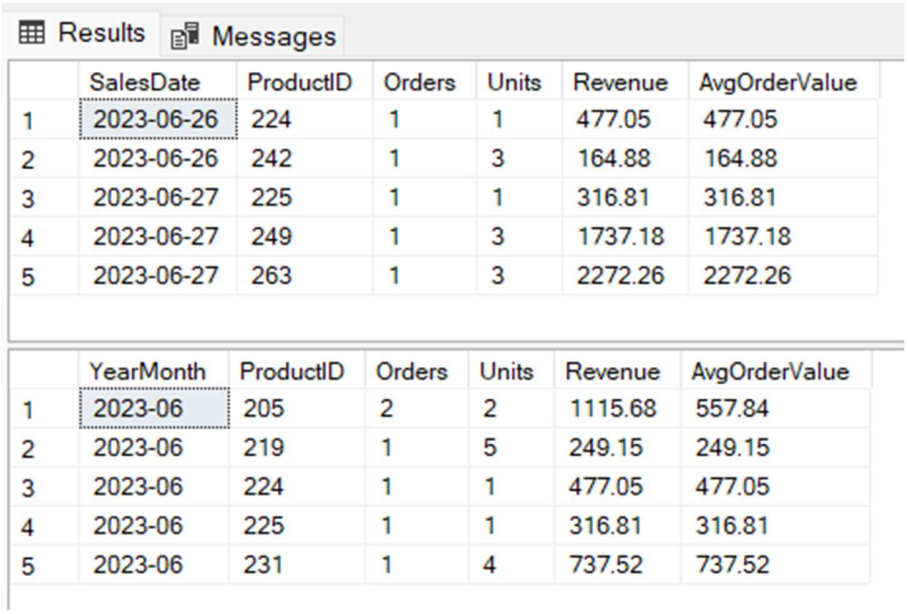
- Daily sales summary

- Monthly sales summary

These tables store key metrics such as:

- Number of orders
- Quantity sold
- Revenue
- Average order value

A SQL view was also created to combine product information, sales results, and review counts. This view is designed to be easily used in Power BI dashboards.



Results Messages

	SalesDate	ProductID	Orders	Units	Revenue	AvgOrderValue
1	2023-06-26	224	1	1	477.05	477.05
2	2023-06-26	242	1	3	164.88	164.88
3	2023-06-27	225	1	1	316.81	316.81
4	2023-06-27	249	1	3	1737.18	1737.18
5	2023-06-27	263	1	3	2272.26	2272.26

	YearMonth	ProductID	Orders	Units	Revenue	AvgOrderValue
1	2023-06	205	2	2	1115.68	557.84
2	2023-06	219	1	5	249.15	249.15
3	2023-06	224	1	1	477.05	477.05
4	2023-06	225	1	1	316.81	316.81
5	2023-06	231	1	4	737.52	737.52

7. Current Status

At this stage:

- The database structure is created
- Data loading is completed
- Aggregated data for reporting is ready

The project is now prepared for visualization and advanced analysis.

Next Steps

The next steps of the project are planned as follows:

1. Build a web user interface to upload datasets (next 2 weeks)

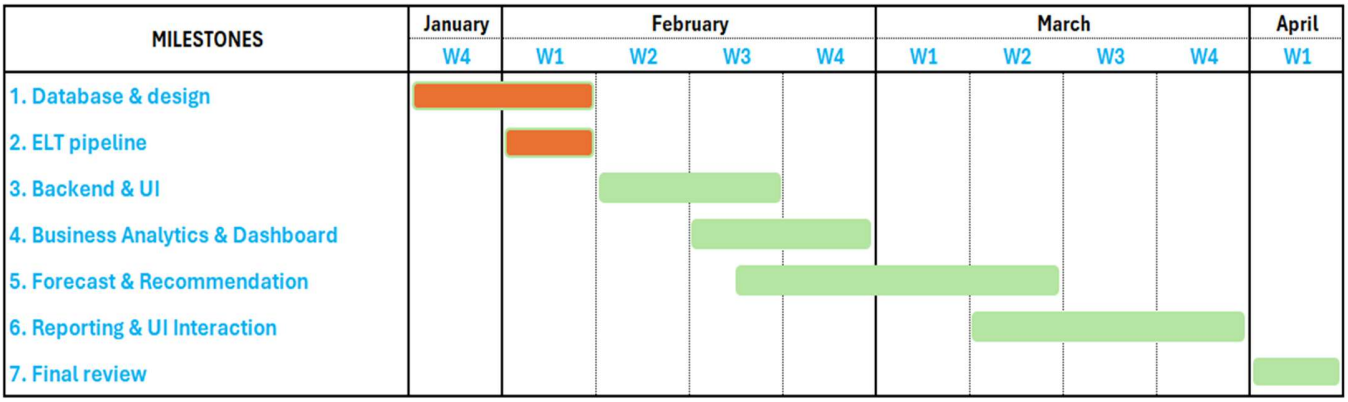
2. Display Power BI dashboards inside the web interface (after step 1)
3. Develop a sales forecasting feature (after step 2)
4. Perform sentiment analysis on customer reviews (after step 3)
5. Write the final project report (after step 4)

Project Planning and Timeline

Time	Milestone	Actions
Jan 26 – Feb 8, 2026	Database & design	• Search for suitable datasets for the project - **DONE** • Identify key business entities, including sales, products, inventory, and customer reviews - **DONE** • Design a relational database schema in SQL Server - **DONE** • Define data relationships and primary keys - **DONE**
Feb 2 – Feb 8, 2026	ELT pipeline	• Implement data ingestion from CSV and Excel files - **DONE** • Load raw data into SQL Server without changing the original content - **DONE** • Check file structure and required fields for correctness - **DONE** • Prepare clean and standardized datasets for analysis - **DONE**
Feb 9 – Feb 22, 2026	Backend & UI	• Develop a Python-based backend for file upload and data processing • Create a simple user interface for uploading data and selecting parameters • Connect the backend to SQL Server for data storage and retrieval

		• Test the complete data flow from file upload to database storage
Feb 16 – Mar 1, 2026	Business Analytics & Dashboard	• Define and calculate key business KPIs such as revenue, sales volume, and average order value • Perform product-level and time-based performance analysis • Build dashboards to show trends and comparisons • Enable time-range filtering for analytics
Feb 19 – Mar 15, 2026	Forecast & Recommendation	• Train forecasting models using historical sales data • Generate future sales and inventory demand forecasts • Compare forecast results with historical performance • Produce basic business recommendations based on analytics and forecasts
Mar 9 – Mar 29, 2026	Reporting & UI Interaction	• Generate automated business review reports • Combine KPIs, trends, forecasts, and recommendations into clear reports • Allow users to select time ranges when generating reports • Present analytical results clearly through the user interface
Mar 30 – Apr 5, 2026	Final review	• Review and validate all system components • Test the accuracy of analytics and report outputs

		• Finalize documentation and the research report • Prepare for project presentation



Work log

Date	Number of Hours	Description
Jan 26, 2026	2	- Searched for datasets that could be used for the project. Websites such as Kaggle and Google Dataset Search were checked. The focus was on finding data with sales information and customer reviews - ChatGPT's prompts: • *"what types of corpus of customer's reviews could be?"* • *"customer's review corpus in csv is considered as structured or unstructured data?"*
Jan 27, 2026	1.5	- Continued searching for datasets on Kaggle, the UCI Machine Learning Repository, and Harvard Dataverse. Different datasets were reviewed to see if they met the project requirements - ChatGPT's prompts:

		• *"review this dataset to see whether is it meet requirements of my project?"* • *"suggest sources of dataset as the project requirements"*
Jan 29, 2026	3	- Looked for more datasets on Data.gov and Kaggle - Researched data scraping techniques from web - ChatGPT's prompts: • *"how to scrap data of sales and client's review from a web"* • *"any legal issues of the scraping?"*
Jan 31, 2026	4	- Selected a dataset from Kaggle due to time limits - Set up environments (VSC, SQL Server Management System, ODBC) - Created IBP database in SQL Server - Created schemas for raw dataset - Researched inserting the raw data into the database by Python - ChatGPT's prompts: • *"which dependencies must be installed for my project?"* • *"how to connect backend and SQL server by python?"* • *"any syntax issue of this table creation?"* • *"how to inject data from csv into SQL server by python?"* • *"explain in step by step these functions"*
Feb 1, 2026	1	- Loaded the raw sales and review data into the database using Python. Checked the results to make sure the data was loaded correctly.

		- ChatGPT's prompts: • *"any wrong syntax or unlogic of this function? and correct it"* • *"explain in details these statements"*
Feb 2, 2026	4	- Studied data cleaning methods using Python. - Created clean tables using dimension and fact tables in SQL Server. - Transformed raw data into clean data - ChatGPT's prompts: • *"why should have dim and fact tables?"* • *"dim & fact are for normalization purpose, is it right?"*
Feb 3, 2026	3	- Studied how to prepare data for Power BI. - Created analysis tables and a summary view in SQL Server. - ChatGPT's prompts: • *"why don't use data directly from the clean tables, instead of creating new analysis tables?"* • *"why used view instead of creating another analysis table? "*
Feb 4, 2026	4	- Drafted the project progress report - ChatGPT's prompts: • *"wording these paragraphs"*
Feb 6, 2026	2	- Updated the project progress report - ChatGPT's prompts: • *"how can use fastapi to connect a html frontend and backend in python"* • *"explain these api in details, line by line"*
Feb 8, 2026	1.5	- Researched front-end's frameworks for uploading the dataset - ChatGPT's prompts:

		• *"which framework works best with html + fastApi?"* • *"compare bootstrap and tailwind"* • *"why should use jinja template? What if not using it?"*