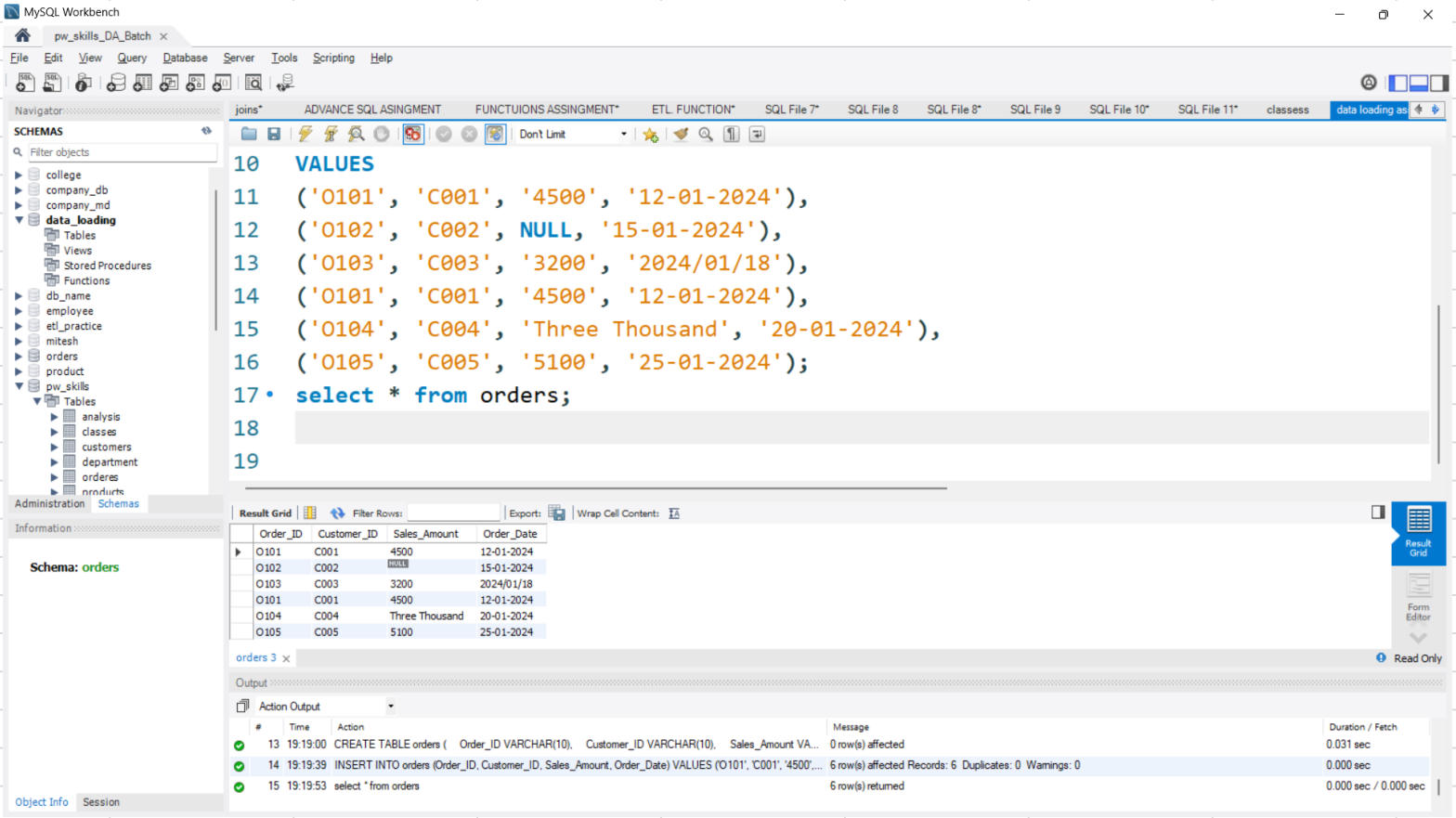
Data Loading (ETL)

DATA SET



QUESTION 1	**Data Understanding** **Identify all data quality issues present in the dataset that can cause problems during data loading?**

ANSWER	**Duplicate Records**	O101 appears twice. Entire row is duplicated.
	Missing Values (NULL)	Sales_Amount for C002 is Null. If column is NOT NULL, data load fails.
	Inconsistent Date Format	Different formats: 12-01-2024 2024/01/18
	Incorrect Data Type (Mixed Data)	Sales_Amount contains: Text value → "Three Thousand" Numeric values (4500, 3200)
	Inconsistent Structure (Header Issue)	Order_ID separated from other columns.
	No Defined Primary Key	If no primary key: Duplicate data possible No uniqueness enforcement

QUESTION 2	Primary Key Validation
	Assum **order_id** s the Primary Key. a) Is the dataset violating the Primary Key rule? b) Which record(s) cause this violation?
ANSWER	Order_ID is the Primary Key
	A Primary Key must:
	Be unique
	Not be NULL
	Is the dataset violating the Primary Key rule?
	Yes, it is violating the Primary Key rule.
	Because a Primary Key must contain unique values, and here we have a duplicate.
	Which record(s) cause this violation?
	the duplicate values is 0101
	Primary key cannot repeat.
	SQL will reject the second O101 during insert if PK is applied.

QUESTION 3		Missing Value Analysis
		Which column(s) contain missing values? a) List the affected records b) Explain why loading these records without handling missing values is risky
ANSWER		
	A)	**The column with missing value is:** **Sales_Amount**
		O102 / C002 / NULL |/15-01-2024
	B)	**NULL values may reduce total revenue.**
		Business may think sales are lower than actual.
		Incorrect KPIs
		If Sales_Amount is used to calculate:
		Total Revenue
		Average Order Value
		Monthly Sales

QUESTION 4

Data Type Validation

Identify records where Sales_Amount violates expected data type rules.
a) Which record(s) will fail numeric validation?
b) What would happen if this dataset is loaded into a SQL table with

ANSWER

A)

The record that violates numeric rules is:

Order_ID	Customer_ID	Sales_Amount	Order_Date
O104	C004	Three Thousand	20-01-2024

"Three Thousand" is text

It cannot be converted into INT or DECIMAL automatically

B)

What would happen if this dataset is loaded into a SQL table with:

Sales_Amount INT

4500 → Inserted

3200 → Inserted

5100 → Inserted

NULL → Inserted

"Three Thousand" → ERROR

QUESTION 5 **Date Format Consistency**

THE ORDERD DATE column has multiple formats.
a) List all date formats present in the dataset
b) Why is this a problem during data loading?

ANSWER

A) **List all date formats present in the dataset**

There are two different date formats:

Format 1 :: DD-MM-YYYY

Examples:

12-01-2024
15-01-2024
20-01-2024
25-01-2024

FORMAT 2 :YYYY/MM/DD

Example:

2024/01/18

B) **Data Type Conversion Failure**

Order_Date DATE

Mixed formats may cause:
- Conversion errors
- Failed insert statements

12 January 2024? (DD-MM-YYYY)

Or December 1, 2024? (MM-DD-YYYY)

QUESTION 6

Load Readiness Decision Based on the dataset condition:
a) Should this dataset be loaded directly into the database? (Yes/No)
b) Justify your answer with at least three reasons

ANSWER

A) Should this dataset be loaded directly into the database?

No, this dataset should NOT be loaded directly into the database.

1) Primary Key Violation (Duplicate Order_ID)

O101 appears twice.
If Order_ID is defined as a PRIMARY KEY, the load will fail.

2) Data Type Issues in Sales_Amount

The value "Three Thousand" is text.
If Sales_Amount is defined as INT or DECIMAL, this record will fail.

3) Missing Value (NULL in Sales_Amount)

One record contains NULL.
If the column is NOT NULL, loading will fail.

4) Inconsistent Date Formats

Two formats exist:
DD-MM-YYYY
YYYY/MM/DD

The dataset should not be loaded directly into the database because it contains duplicate primary key values, data type inconsistencies in Sales_Amount, missing values, and inconsistent date formats.

QUESTION 7

Pre-Load Validation Checklist
List the exact pre-load validation checks you would perform on this dataset before loading.

ANSWER

1) Primary Key Validation (Uniqueness Check)

Check for duplicate Order_ID

Ensure no NULL values in Order_ID

Confirm it can safely be defined as PRIMARY KEY

2) Missing Value Check

Identify NULL values in:

Sales_Amount

Order_Date

Customer_ID

3) Data Type Validation

Sales_Amount

Must contain only numeric values

No text like "Three Thousand

4) Date Format Standardization

Ensure all dates follow a single format (preferably YYYY-MM-DD)

Validate date correctness (no invalid dates like 32-01-2024)

5) Referential Integrity Check

Validate that Customer_ID exists in Customer Master table

QUESTION 8	**. Cleaning Strategy** **Describe the step-by-step cleaning actions required to make this dataset load-ready.**
ANSWER	**1: Remove Duplicate Records**
	Problem: Order_ID = O101 appears twice.
	Identify duplicate Order_ID
	Keep only one valid record
	Delete or archive duplicate row
	2: Handle Missing Values (Sales_Amount)
	Problem: One record has NULL.
	Possible Actions (Based on Business Rule):
	Replace with actual value (if available)
	Impute with 0 (if business allows)
	Mark as “Pending” and exclude from repor
	3: Fix Data Type Issues (Sales_Amount)
	Problem: "Three Thousand" is text.
	Convert "Three Thousand" → 3000
	Ensure entire column is numeric
	Validate using numeric check
	4: Standardize Date Format
	12-01-2024
	2024/01/18

		5: Validate Column Constraints						
		Set Order_ID → PRIMARY KEY						
		Set Sales_Amount → NOT NULL (if required)						
		Set Customer_ID → FOREIGN KEY (if customer table exists)						
		6: Trim & Standardize Text Fields						
		Remove extra spaces						
		Ensure consistent case formatting						
		Check for hidden characters						
		7: Business Rule Validation						
		Ensure Sales_Amount > 0						
		Ensure no future dates (if rule applies)						
		Ensure each order belongs to valid customer						

QUESTION 9

Loading Strategy Selection Assume this dataset represents daily sales data.
a) Should a Full Load or Incremental Load be used?
b) Justify your choice

ANSWER

Incremental Load should be used.

Entire historical data will reload every day.
It wastes time and system resources.

Incremental load:
Loads only new records for the day.
Much faster and more efficient.

Performance & Scalability

As data grows:

Full Load becomes slow.
Database locking may increase.
ETL window may exceed allowed time.

Incremental load:

Handles large datasets better.
Scales easily over time.

Reduces Risk of Duplicate Data

With proper incremental logic (using Order_Date or Order_ID):

		Only new records are inserted.							
		Existing records are not reloa							

QUESTION 10

BI Impact Scenario Assume this dataset was loaded without cleaning and connected to a BI dashboard.
a) What incorrect results might appear in Total Sales KPI?
b) Which records specifically would cause misleading insights?
c) Why would BI tools not detect these issues automatically?

ANSWER

1) Inflated Total Sales

O101 is duplicated
Sales amount 4500 will be counted twice
So revenue will be higher than actual.

2) Underestimated Total Sales

One record has NULL Sales_Amount
NULL is ignored in SUM() calculation

3) Missing Sales Value

"Three Thousand" is text

BI tool may:

Ignore it
Treat it as 0

c) Why would BI tools not detect these issues automatically?

BI Tools Assume Data is Clean

BI tools (like dashboards) focus on:
Visualization
Aggregation
Reporting

Duplicate Data Looks Legitimate

Two identical rows look valid.
BI tool cannot know:
It is duplicate
Or a real repeated transaction