Informatica Hands-On Challenge: Super_Store Analysis

- **Introduction:** You are provided with a sample dataset from a retail store, Super_Store. This dataset contains information about orders, customers, products, and sales. Your task involves cleaning the data, analyzing sales, customer orders, customer geography, and order processing time using Informatica PowerCenter.

Data Preparation:

- **Oracle SQL Setup:**
- Log in to Oracle SQL Developer in **Admin** connection using the credentials:
- **Username:** system
- **Password:** Admin

Create a table named **Super_Store** with the provided structure

Row_ID	INT
Order_Date	DATE
Ship_Date	DATE
Ship_Mode	VARCHAR(50)
Customer_ID	VARCHAR(50)
Customer_Name	VARCHAR(50)
Segment	VARCHAR(50)
Country	VARCHAR(50)
City	VARCHAR(50)
State	VARCHAR(50)
Postal_Code	VARCHAR(50)
Region	VARCHAR(50)

Product_ID	VARCHAR(250)
Category	VARCHAR(250)
Sub_Category	VARCHAR(250)
Product_Name	VARCHAR(250)
Sales	INT

- **NOTE** : while loading data into table update **order_date, ship_date** Date Format to **DD/MM/YYYY**

Load **superstore_data.csv** into the **Super_Store** table.

You are given data set is in the **"~\Desktop\Project\miniproject-informaticasuper_store \"**

Informatica Repository Setup:

Connect to the Informatica repository manager using the following credentials:

- **Username:** Administrator
- **Password:** Administrator

Create a folder named **Super_Store** in the repository manager.

How to Import Source Table in Source Analyzer

Following are the steps to import source table in Informatica Source Analyzer:

Step 1) Go to "Sources" option In

source analyzer

1. Click on tab "Sources" from the main menu
2. Select import from database option, after this ODBC Connection box will open.

Step 2) Create ODBC connection

- We will now create ODBC connection

1. Click on the button next to ODBC data Source(...).
2. On the next page, Select user DSN tab and click Add button.
3. Select oracle wire protocol
4. On the next page, select the general tab and enter the database details. Then click connect.

- **Data Source name** : oracle
- **Host** : localhost
- **port** : 1521
- **sid** : xe

Create Connections for Workflow Manager

To Create a Relational Connection

Step 1: In Workflow Manager

- Click on the **Connection** menu
- Select **Relational** Option

Step 2: In the pop up window

- Select **Oracle** in type
- Click on the **new** button

Step 3: In the new window of connection object definition

- Enter Connection Name (oracle)
- Enter **username - system**
- Enter **password – Admin**
- Enter **connection string - xe**
- Leave other settings as default and Select OK button **Note :** For more **credentials**, like for designer, knidly check in the **Readme** File.

Note : Please Follow the naming conventions in the problem statement

Data Cleaning:

- **Mapping Name:** Map_Cleaned_Data

- **Workflow Name:** Workflow_Cleaned_Data
- **Session Name:** Session_Cleaned_Data
- **Target Table:** Super_Store_Cleaned_Data

Operations:

- Remove duplicates from the dataset to ensure data integrity.
- Filter records where Country is 'United States' to focus on domestic orders.
- Extract numeric part from Customer_ID to standardize customer identification.(EX: CH-1234, extract 1234)
- Concatenate Customer_ID and Customer_Name with '-' to create a unique identifier for each customer.(Ex: 1234-Charlies, Extracted_ID-Customer_name) and store it in Customer_Id_Name Column
- Drop the customer_id, Customer_name Columns
- After cleaning Load data into the **Super_Store_Cleaned_Data** target table (For columns check sample output)
- **Sample Output :** Reamining columns and additional with **'CUSTOMER_ID_NAME'** COLUMN.

CUSTOMER_ID_NAME
21925-Zushuss Donatelli
16585-Ken Black
21520-Tracy Blumstein

NOTE : Super_Store_Cleaned_Data table data is used for the below every tasks.

Analysis Tasks:

Task 1: Sales Summary

- **Mapping Name:** Map_Sales_Summary
- **Workflow Name:** Workflow_Sales_Summary
- **Session Name:** Session_Sales_Summary
- **Target Table:** Sales_Summary

Problem Statement: Summarize total sales and average sales for each customer. Identify customers with significant contribution to overall sales.

Operations:

- Filter the records region in East and state in New York to focus on a specific customer base.
- Convert the sales amount from USD to INR (conversion rate= 84) and store it in AMOUNT column.
- Calculate the interest rate for amount values greater than 5000 (interest rate is 10%) & store in new column INTREST. Sum up the amount and interest to SALES column.
- Calculate the total sales and average sales for each customer. Filter customers with total sales greater than 5000 and average sales greater than 500 to focus on significant contributors.
- Drop the unnecessary columns, kindly check the sample output.
- Load data into the **Sales_Summary** target table (For columns check sample output)
- After completing of mapping, in the workflow manager.

Sample Output:

CUSTOMER_ID_NAME	TOTAL_SALES	AVG_SALES
10060-Adam Bellavance	409458	81892
17470-Mark Packer	154535	25756
14815-Harold Pawlan	121136	60568

Task 2: Customer Order Analysis

- **Mapping Name:** Map_Order_Analysis
- **Workflow Name:** Workflow_Order_Analysis
- **Session Name:** Session_Order_Analysis
- **Target Table:** Order_Analysis

Problem Statement: Analyze customer orders to determine the most frequent buyers and their order patterns.

Operations:

- Filter records for customers in category 'Furniture' and City in 'New York City' to analyze local customer behavior.
- Create new column orders_count, Calculate the count of orders for each customer to determine their order frequency.
- Categorize orders based on the number of orders. Orders are less than 10 then 'Low', orders are between 10-20 then 'Medium', orders are greater than 20 then 'High'.

- Sort the results by order count in descending order to identify the most frequent buyers and get only top 8 records.
- Generate a unique number for each row into column Sno. values should start from 11.
- Drop the unnecessary columns, kindly check the sample output.
- Load data into the **Order_Analysis** target table (For columns check sample output)

Sample output:

SNO	CUSTOMER_ID_NAME	ORDERS_COUNT	ORDERS_CATEGORY
11	18355-Nat Gilpin	3	Low
12	12805-Cynthia Voltz	2	Low
13	16435-Katrina Willman	2	Low
14	17470-Mark Packer	2	Low
15	10225-Alan Schoenberger	1	Low

Task 3: Customer Geography Analysis

- **Mapping Name:** Map_Geography_Analysis
- **Workflow Name:** Workflow_Geography_Analysis
- **Session Name:** Session_Geography_Analysis
- **Target Table:** Geography_Analysis

Problem Statement: Analyze customer distribution across different regions to identify potential market segments.

Operations:

- Filter records for customers in Segment 'Consumer'.
- Combine column customer_id_name and region using ' _' and place them in new column customer_region
- Store the output in respective tables based on the region.
- Drop the unnecessary columns, kindly check the sample output.
- Load data into the **EAST_CUSTOMER_BASE, WEST_CUSTOMER_BASE, SOUTH_CUSTOMER_BASE** target table (For columns check sample output)

Sample output: EAST_CUSTOMER_BASE

CUSTOMER_REGION	PINCODE	STATE	CATEGORY
19960-Ryan Crowe_East	43229	Ohio	Office Supplies
19960-Ryan Crowe_East	43229	Ohio	Office Supplies
20725-Steven Cartwright_East	19805	Delaware	Office Supplies

Sample output: WEST_CUSTOMER_BASE

CUSTOMER_REGION	PINCODE	STATE	CATEGORY
16885-Lena Creighton_West	95661	California	Office Supplies
12130-Chad Sievert_West	90004	California	Office Supplies
11710-Brosina Hoffman_West	90032	California	Furniture

Sample output: SOUTH_CUSTOMER_BASE

CUSTOMER_REGION	PINCODE	STATE	CATEGORY
16270-Karen Daniels_South	22153	Virginia	Office Supplies
18385-Natalie Fritzler_South	39212	Mississippi	Furniture
19780-Rose OBrian_South	38109	Tennessee	Furniture

Task 4: Order Processing Time Analysis

- **Mapping Name:** Map_Order_Processing
- **Workflow Name:** Workflow_Order_Processing
- **Session Name:** Session_Order_Processing

- **Target Table:** Order_Processing

Problem Statement: Evaluate order processing efficiency by analyzing the time taken between order placement and shipment,

Operations:

- Calculate the repeated orders for each product subcategory and store them in ORDERS_COUNT column.

- Categorize the repeat orders (e.g., less than 10 orders Low Sales, between 10-30 Average Sales, more than 30 orders Best sales).

- Count the number of orders falling with in each category to analyze product sales. Load the data into REPEAT_ORDERS table.

- Calculate the processing days for each order by finding the difference between order date and ship date and store it in new column Processing_days.

- Categorize processing days (e.g., Less than 1 day then One-Day Delivery, 1 to 2 days then Two-Day Delivery, 3 or more days then Standard Delivery).

- Count the number of orders falling with in each categorize processing days for each to analyze processing days distributions. Load the data into **Order_Processing** table.

- Drop the unnecessary columns, kindly check the sample output.

- Load data into the **Order_Processing, REPEAT_ORDER** target tables (For columns check sample output)

Sample Output:

CATEGORISE_PROCESSING_DAYS	ORDERS_COUNT
One-Day Delivery	17
Standard Delivery	765

Two-Day Delivery	**208**

SALES_CATEGORY	PRODUCT_SUB_CATEGORY_COUNT
Average Sales	**4**
Best Sales	**4**
Low Sales	**9**