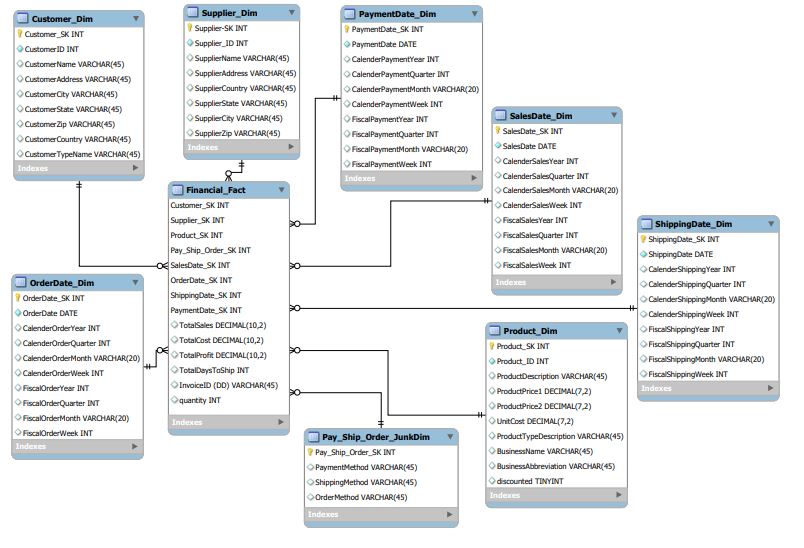
**ISTE 724: DATA WAREHOUSING**

**LAB 3, Team 2**

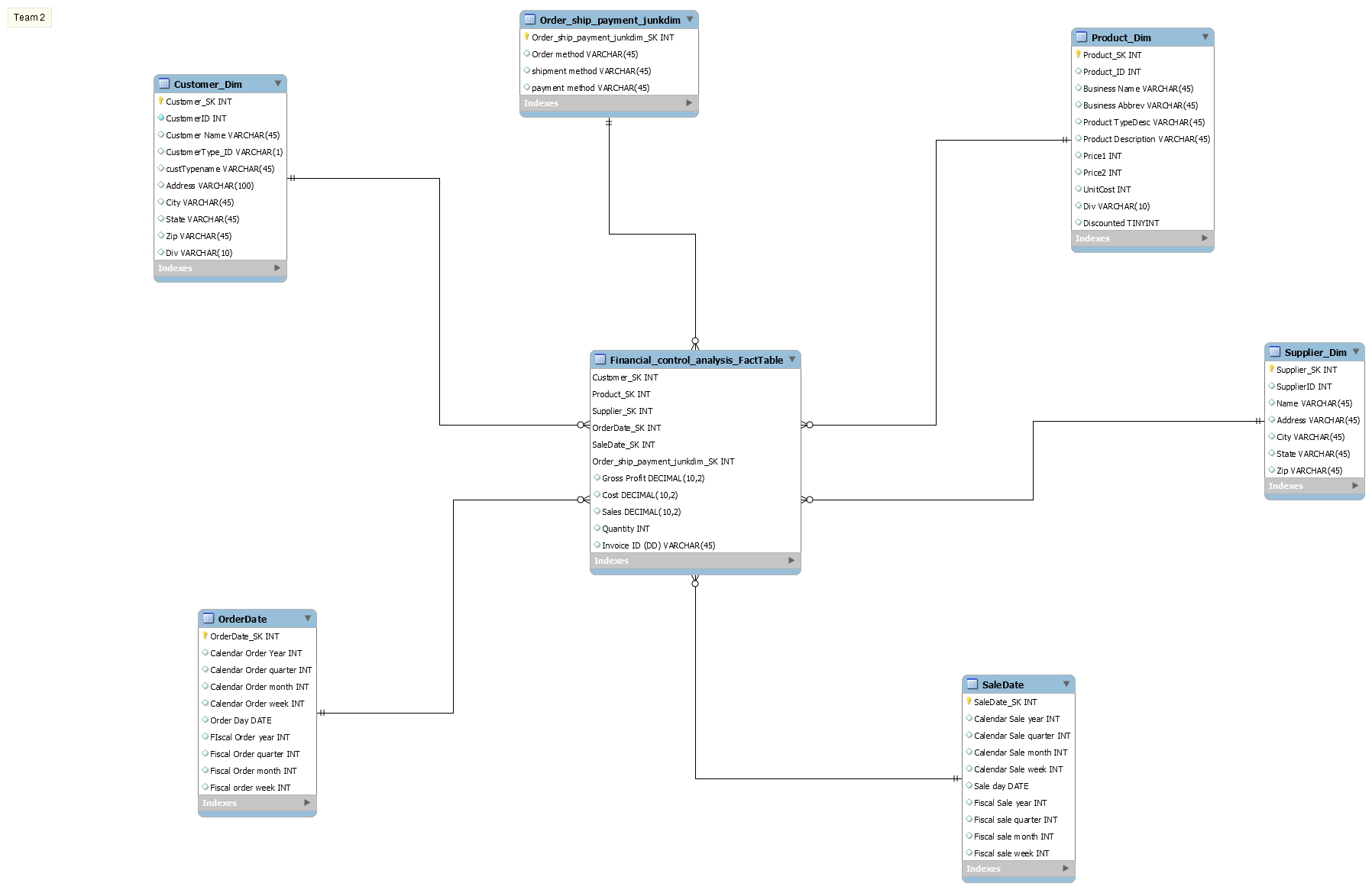
**RATIONAL, QUERIES AND DATA STAGING ACTIVITIES**

**LAB 2, Dimensional Model:**

****

In lab2, we had PaymentDate\_Dim and ShippingDate\_Dim but we do not have any information or data related to this dimensions and hence, we have removed this from the dimensional model. Another addition to our dimensional model is that we have included a new attribute called ‘DIV’ in the customer table that will help segregate that which customer is associated or has had a transaction related to which division. This will ease the analysis for the company as they can view customer details and much more based on specific divisions across the company. The next change is that we have added a ‘Sales By’ attribute in the product dimension that will indicate that which division’s product that product belongs to that will eventually help the company to keep track of their equipment’s being used across the divisions. Finally, a ‘DIV’ field is also added to the fact table that will help the company to track sales, profit, costs as per individual divisions. We also removed the country dimension from the address part for the customer and supplier as all data is associated to US or a specific company in the US. Finally, we have also cleaned, transformed the company data as per the rules specified and loaded it in the following dimension.

**FINAL AGREED UPON DIMENSIONAL MODEL:**

****

**Appendix A: ETL STAGING ACTIVITIES**

|  |  |  |  |
| --- | --- | --- | --- |
| **FILE** | **ATTRIBUTE** | **PROBLEM** | **RESOLUTION STRATEGY** |
| **PEC Customer** | **Name, Address, City, State, ZIP, CustTypeID, CustID** | They had inverted commas | Removed by using replace in string. |
|  | **Address, CustTypeID** | Had additional commas, period signs | Removed them using replace in string |
|  | **Address** | Had continuous numbers | Addes commas between 2 separating numbers of number entities using regex |
|  | **Address** | St, Ave, Rd, Av were present | Changed to Street, Avenue, Road, Avenue using regex |
|  | **Name** | Inc., Corp, Co. were present | Changed ti Incorporated. ,Corporation and Company sing regex. |
|  | **DIV** | Column was not present | Added column with value as PEC using value mapper. |
|  | **CustTypeName** | CustTypeID missing in customer dimension | Added this using stream lookup |

|  |  |  |  |
| --- | --- | --- | --- |
| **FILE** | **ATTRIBUTE** | **PROBLEM** | **RESOLUTION STRATEGY** |
| **TPCW Customer** | **ALL** | Values being shown in multiple columns in source csv | Made extra fields to access those additional columns.  Then concatenated all fields to bring them into one column with a ‘;’ as delimiter. |
|  | **ALL** | All values brought under one column | Separated them using the split fields with ‘;’ as delimiter and getting all columns as separate columns. |
|  | **ALL** | All fields have inverted commas | Removed them using replace in string |
|  | **Address** | Had ‘.’ And ‘,’ | Removed them using replace in string to maintain uniformity with PEC |
|  | **DIV** | Missing column | Added new column with TPCW as value |
|  | **ZIP** | Missing values | Replaced with N/A using IF Field value is null |
|  | **CustTypeID** | Missing Values | Replaced with N/A using IF Field value is null |
|  | **Name** | Inc., Corp, and Co., were present | Replaced with Incorporated. , Corporation and Company as done in PEC. |
|  | **ZIP** | Length was less than 5 | Replaced length with 5 digits by adding trailing zeros using javascript. |
|  | **State** | State was full name | We have maintained shortened versions of states across all divisions using javascript. |
|  | **CustTypeName** | Missing column | Added this using stream lookup with CustomerType |

|  |  |  |  |
| --- | --- | --- | --- |
| **FILE** | **ATTRIBUTE** | **PROBLEM** | **RESOLUTION STRATEGY** |
| **TPCE Customer** | **Address 1 and Address2** | Two separate fields | Concatenated into one Address field with address1 then address 2. |
|  | **Address** | Alignment issues | Left aligned all values in column |
|  | **DIV** | Missing column | Added column with TPCE as value |
|  | **Address 1, Address 2** | Need to be removed as new Address field created | Removed using select function |
|  | **Address** | 2 numbers were together making difficult to interpret | Added a comma separating the 2 fields to make it easier to read. |
|  | **Name** | Inc., Corp and Co., was present | Replaced with Incorporated. , Corporation and Company as maintained in others. |
|  | **CustTypeName** | Missing column | Added column using stream lookup at customer type table |

|  |  |  |  |
| --- | --- | --- | --- |
| **FILE** | **ATTRIBUTE** | **PROBLEM** | **RESOLUTION STRATEGY** |
| **TPCE CustomerType (Used only this and cleaned only one as the other customertype tables were same. Only this had a little anomaly, and hence cleaned this.)** | **CusttypeID** | Enclosed in double inverted commas | Removed using replace in string |
|  | **TypeName** | Enclosed in double inverted commas | Removed using replace in string |
|  | **TypeName** | Presence of \_ instead of / in other customerTypes | Replaced using replace in string |

\***The surrogate key for customer was added after combining all the cleaned customer files into one cleaned customer text file using the add sequence function. Then that was loaded in mysql workbench. It wasn’t added individually in each customer file.**

|  |  |  |  |
| --- | --- | --- | --- |
| **FILE** | **ATTRIBUTE** | **PROBLEM** | **RESOLUTION STRATEGY** |
| **TPCE Supplier(Created a new supplier dimension)** | **ADDR1** | Attn : is present which is not supposed to be there in address | Removed using replace in string. |
|  | **ADDR1, ADDR2** | They both are separate and as we have done in customer table, we maintain a similar format under 1 address column | Concatenated Addr1 and Addr2 |
|  | **Name** | TPCE, TPCW and PEC as suppliers are missing | Used append streams and filter rows function to filter out TPCE and TPCW from PEC customer table and PEC from TPCE customer table |
|  | **Address** | Multiple spaces not present in address column to separate words | Added a space using replace in string option |
|  | **Name** | Inc., and Corp were present | Changed to Incorporated. , and Corporation |
|  | **SUPPID** | The new id’s were not unique or were out of place | Added a new auto increment column that gave all suppliers a unique value |
|  | **ID** | Redundant column as SUPPID was added | Removed using the select function |
|  | **Supplier\_SK** | Missing surrogate key column | Added column using add sequence function |

|  |  |  |  |
| --- | --- | --- | --- |
| **FILE** | **ATTRIBUTE** | **PROBLEM** | **RESOLUTION STRATEGY** |
| **PEC Invoice** | **All** | Shifted entries for two rows | Shifted manually for all columns of those two rows in the csv file. |
|  | **ShipMethod** | Aiir, train, trran, trrain, trick, tuck, n/a as field values are present | Corrected to Air, Train, Truck and N/A respectively. |
|  | **SalesDate** | * As separator between dates | Standardised MM/DD/YY as format across all dates. |
|  | **OrderDate** | * As separator between dates | Standardised MM/DD/YY as format across all dates. |
|  | **SalesDate** | Was in string format initially | Changed to date format and in format mentioned above |
|  | **OrderDate** | Was in string format initially | Changed to date format and in format mentioned above |

|  |  |  |  |
| --- | --- | --- | --- |
| **FILE** | **ATTRIBUTE** | **PROBLEM** | **RESOLUTION STRATEGY** |
| **TPCW Invoice** | **InvoiceID** | Bad record | Removed using filter rows for invoiceID 26511 |
|  | **CustID** | Negative CustID | First changed datatyoe from String to integer and then using Javascript, removed negatives and replaced with positive value. |
|  | **SalesDate** | Inconsistencies between date format | All dates were changed to MM/DD/YY format |
|  | **Quantity** | Presence of null values | Replaced with 0 |
|  | **Discounted** | Presence of null values | Replaced with 0 |

|  |  |  |  |
| --- | --- | --- | --- |
| **FILE** | **ATTRIBUTE** | **PROBLEM** | **RESOLUTION STRATEGY** |
| **PEC Product** | **PRODID, prodDescription, supplierName, and ProductTypeID** | Enclosures present | Removed enclosures using replace in string method |
|  | **prodDescription** | Equip was present | Standardised as Equipment across all departments. |
|  | **UnitCost** | Missing column in product table | Added this column using stream lookup in PEC Product with PEC Manufacturing and also calculated unitcost. UnitCost was calculated by grouping Year, month and prodID of PEC invoice and PEC manufacturing. Then merge joined invoice and manufacturing and finally Using calculator, calculated unitcost using manufacturing cost/quantity for that month. |
|  | **SupplierName** | Values were null | Replaced with N/A. |
|  | **Business Name, Abbrev, BUID, ProdtypeID, typedesc** | Missing fields | Added using PEC Product and TPCW Business Unit by stream lookup on ProdTypeID |
|  | **Supplier address, city, state and zip** | Missing fields | Added these columns using stream lookup between PEC product and Supplier on Supplier Name. |
|  | **Discounted** | Missing fields | Using stream lookup on PEC product and PEC Invoice on productID |
|  | **SalesBY** | Missing field | Added additional field with value PEC |
| **TPCW Business Unit (Used this table for lookup for getting BUID, Business Name and Abbrev. It is same across all 3 units)** |  |  |  |
|  | **Abbrev** | Missing abbrev for business name Miscellaneous | Added Misc as Abbrev using modified javascript |
|  | **Business Name, Abbreviation** | Missing Fields in PEC Product | Got using stream lookup on BUID between TPCW Product Type and Business Unit. |
| **TPCW Product** |  |  |  |
|  | **PRODID, prodDescription, supplierName, and ProductTypeID, ADDR1, ADDR2, CITY, ZIP, PRICE1, PRICE2, UNITCOST** | Enclosures present | Removed enclosures using replace in string method |
|  | **ADDR1** | Presence of Attn : | Removed using replace in string |
|  | **Proddesc** | Equip was present | Standardised as Equipment across all departments. |
|  | **CITY, STATE** | Were one field in table | Split based on ‘,’ to separate the two columns |
|  | **ADDR1. ADDR2** | Separate fields | Concatenated into 1 as we have done for all addresses across all departments |
|  | **ADDR1. ADDR2** | Additional as new address field added | Removed using the select function |
|  | **SUPPNAME** | Inc., corp., and co were there | Standardised to Incorporated. , corporation and company. |
|  | **Address** | St., Ave., Rd., Av., were present | Standardised to Street, Avenue and Road. |
|  | **Business Name, Abbrev, BUID, typedesc** | Missing fields | Added using TPCW Product and TPCW Business Unit by stream lookup on ProdTypeID |
|  | **SalesBY** | Missing field | Added additional field with value TPCW |
|  | **Discounted** | Missing fields | Using stream lookup on TPCW product and TPCW Invoice on productID |
| **TPCE Product** | **PRODID, prodDescription, supplierID, and ProductTypeID,**  **ProdTypeDescription, PRICE1, PRICE2, UNITCOST** | Presence of enclosures | Removed Enclosures using replace in string. |
|  | **ProductTypeDescription** | Equip was present | Standardised as Equipment across all departments. |
|  | **ProdTypeID** | Presence of trailing 0’s | Removed using regex |
|  | **SupplierName, Address, City, Zip, State** | Missing for supplier | Added using stream lookup on supplierID of supplier table in TPCE and this table |
|  | **SalesBY** | Missing field | Added additional field with value TPCE |
|  | **Address** | Presence of Attn : | Removed using replace in string |
|  | **Discounted** | Missing fields | Using stream lookup on TPCE product and TPCE Invoice on productID |
|  | **TypeDesc, BUID, Business Name and Abbrev** | Missing Fields | Stream lookup on ProdTypeId with Business Unit and Product Type |

|  |  |  |  |
| --- | --- | --- | --- |
| **File (.csv & others if applicable)** | **Attribute** | **Problem** | **Resolution Strategy** |
| **PECInvoice.csv**  **(Junk Dimention)** | Invoice, custID, salesDate, prodid, amt, qty, shipMethod, shipCost, paymentMethod, orderMethod, orderDate, discounted | shipMethod, paymentMethod, orderMethod are junk attributes, part of an invoice dimension. So, we segregated them out and put that in junk dimension. | Since the cardinality of these attributes is very low (4, 4, 3), we generated the combinations of these values and populated the table rows with it. Hence, there are in total 4\*4\*3 i.e. 48 unique rows. |
|  |  | **shipMethod**   * There are many variations of air, truck, and train values. * Comprises invalid entries. | **shipMethod**   * Converted invalid entries to N/A * Converted values’ variations back to original form * Variations were :(aiir, trainn, trran, trrain, tuck, trick) * Values entered in wrong column hence shifted values in excel file. |
|  |  | **PaymentMethod**   * Comprises invalid entries. | **PaymentMethod**   * Values entered in wrong column hence shifted values in excel file. |
|  |  | **OrderMethod**   * Comprises invalid entries. | **OrderMethod**   * Values entered in wrong column hence shifted values in excel file. |

|  |  |  |  |
| --- | --- | --- | --- |
| **File (.csv & others if applicable)** | **Attribute** | **Problem** | **Resolution Strategy** |
| **PECInvoice.csv**  **(Date Dimension)** | Invoice, custID, salesDate, prodid, amt, qty, shipMethod, shipCost, paymentMethod, orderMethod, orderDate, discounted | **salesDate**   * Invalid entries in rows * Incorrect date formats * Missing fiscal fields. * Missing separate month, week and quarter fields. | **salesDate**:   * Deleted rows comprise all invalid entries * Changed date format from (MM-dd-YY) to (MM/dd/YY) * Converted non-dates entries to 12/31/99 * Generated fiscal fields (year, quarter, month, week) * Generated calendar fields (year, quarter, month, week) from given date field |
|  |  | **orderDate**   * Invalid entries in rows. * Incorrect date formats * Blank fields * Missing fiscal fields. * Missing separate month, week and quarter fields. | **orderDate**:   * Deleted rows that comprises all invalid entries * Changed date format from (MM-dd-YY) to (MM/dd/YY) * Converted non-dates and blank entries to 12/31/99 * Generated fiscal fields (year, quarter, month, week) * Generated calendar fields (year, quarter, month, week) from given date field |
| **TPCWInvoice.csv**  **(Date Dimension)** | Invoice, custID, prodID, salesDate, amt, qty, discounted | **salesDate**:   * Invalid date entries * Incorrect date formats * Invalid year format * Missing fiscal fields. * Missing separate month, week and quarter fields. | **salesDate**:   * Deleted a row that comprises all invalid entries * Changed date format from (dd-MM-YY) to (MM/dd/YY) * Changed day and month formats of given date. E.g. 4/3/05 to 04/03/05 to follow same format. * Converted 20-08-05372 entry to 08/20/05 * Generated fiscal fields (year, quarter, month, week) * Generated calendar fields (year, quarter, month, week) from given date field |
| **TPCEInvoice.csv**  **(Date Dimension)** | InvoiceID, custID, salesDate | ---- | * All columns are already cleaned except changed day and month formats of given date. E.g. 4/3/05 to 04/03/05 to follow same format. |

**DATA TRANSFORMATION:**

|  |  |
| --- | --- |
| **DM TABLE** | **IMAGE CREATION PROCESS** |
| **Customer Dimension** | 1. Read files for TPCE, TPCW and PEC Customers using text file input. 2. Cleaned files for enclosures, extra commas, maintained uniformity across 3 divisions for road, avenue, street, corporation, company and incorporated. 3. Handled null values by replacing with N/A, removed or deleted dirty data. 4. Used stream lookup with customer type to retrieve customer types using custTypeID. 5. Sorted the customer data after putting into 1 file on customer name. 6. Added surrogate key auto increment column and finally loaded to Customer Dimension. |
| **Supplier Dimension** | 1. Input taken from TPCE 2. 3 values for TPCE, TPCW and PEC were missing in TPCE file. 3. Used TPCE and PEC along with filter rows to fetch PEC, TPCE and TPCW data for suppliers. 4. Used append streams to put information into one table. 5. Added additional unique values for the 3 suppliers using add sequence and removing old id’s that was also the surrogate key column. 6. Added a unique function to check for duplicates for entire row. 7. Finally, loaded into Supplier Dimension. |
| **Junk Dimension** | 1. Took input from invoice data from PEC. 2. Split 3 select methods, each for shipmethod, paymentmethod and ordermethod. 3. Standardised names in each of the attributes for ship, order and payment methods. 4. Sorted each attribute 5. Added a unique constraint 6. Added N/A to payment and order. 7. Performed a Cartesian product on all 3 attributes to get all possible combinations of the 3 attributes. 8. Added surrogate key column using add sequence. 9. Finally, loaded this in the Order\_Pay\_Ship Junk dimension. |
| **Sale Dimension** | 1. Took invoice inputs from all 3 divisions. 2. Changed date type from string to date. 3. Then standardised the format to be as MM/DD/YY. 4. Created new columns for year, quarter, month, week for calendar. 5. Calculated fiscal using javascript. 6. Merged all the 3 in one table. 7. Sorted on salesdate 8. Added surrogate key column using add sequence. 9. Finally added the data in SalesDate Dimension. |
| **Order Dimension** | 1. Used data from PEC invoice. 2. Changed date type from string to date. 3. Then standardised the format to be as MM/DD/YY. 4. Created new columns for year, quarter, month, week for calendar. 5. Calculated fiscal using javascript. 6. Merged all the 3 in one table. 7. Sorted on salesdate 8. Added surrogate key column using add sequence. 9. Finally added the data in SalesDate Dimension. |
| **Product Dimension** | 1. Used all 3 divisions product table. 2. Cleansed data for each. 3. Calculated unit cost for PEC product as they were missing using manufacturing table and PEC invoice. 4. For all 3 divisions, looked up with Product type and business unit to get business name, abbrecv and buid. 5. Finally lookeup up for supplier data as well to get all supplier level details in tables. 6. Added a surrogate key column using 7. Added in a single table and added to Product Dimension and sorted on Product Description. |
| **Invoice Data** | 1. Took input from PEC and TPCW 2. Removed bad Id’s as it contained incorrect data. 3. Vales were standardised, dates were changed from string to date. 4. Removed shipCost attribute 5. Stream lookup done with PEC Product to get price and unitcost. 6. Used javascript to calculate gross profit, and total sales. 7. Null values replaced with N/A. 8. For TPCE, we looked up using TPCE invoice, and updated in TPCE invoice detail. 9. Then looked up again from TPCE product to get price and unitcost. 10. Used javascript to calculate gross profit, and total sales. 11. Joined all of the invoices, sorted on invoiceID and finally loaded into a text file. |

**TABLE POPULATION:**

|  |  |
| --- | --- |
| **DM TABLE** | **TABLE POPULATION PROCESS** |
| **FACT TABLE LOAD PROCESS** | 1. First too into consideration the invoice cleansed table. 2. Then we loaded salesdate dimension using table input. 3. Then used stream lookup on salesdate and Invoice to get saledate surrogate key. 4. Then used stream lookup on orderdate with previous lookup output to get orderdate surrogate key. 5. Then used stream lookup on customer dimension with previous lookup output on custmerID to get customer surrogate key. 6. Then used stream lookup on product dimension with previous lookup output on productID to get product surrogate key. 7. Used product dimension and supplier dimension lookup on supplier name to get supplier sk and finally this lookup and previous lookup output on prodID to get final supplier surrogate key value. 8. For Junk, we concatenated the 3 attribute values from previous step output and also in the junk dimension individually. 9. Then performed a lookup on these concatenated values to retrieve the junk surrogate key. 10. Then we handled the null values in surrogate keys with a value 9999999. 11. Sorted on all the surrogate keys. 12. Performed the unique function on all the surrogate keys. 13. Finally loaded this in the fact table. |

**\*We have added mysql scripts to add 1 record in each of the dimensions. The values were 9999999 for surrogate keys and 9999999 for respective ID’s and nulls to the rest of the attributes. This was needed so as to have a mapping between the fact and the dimensions for the corrected null surrogate keys.**

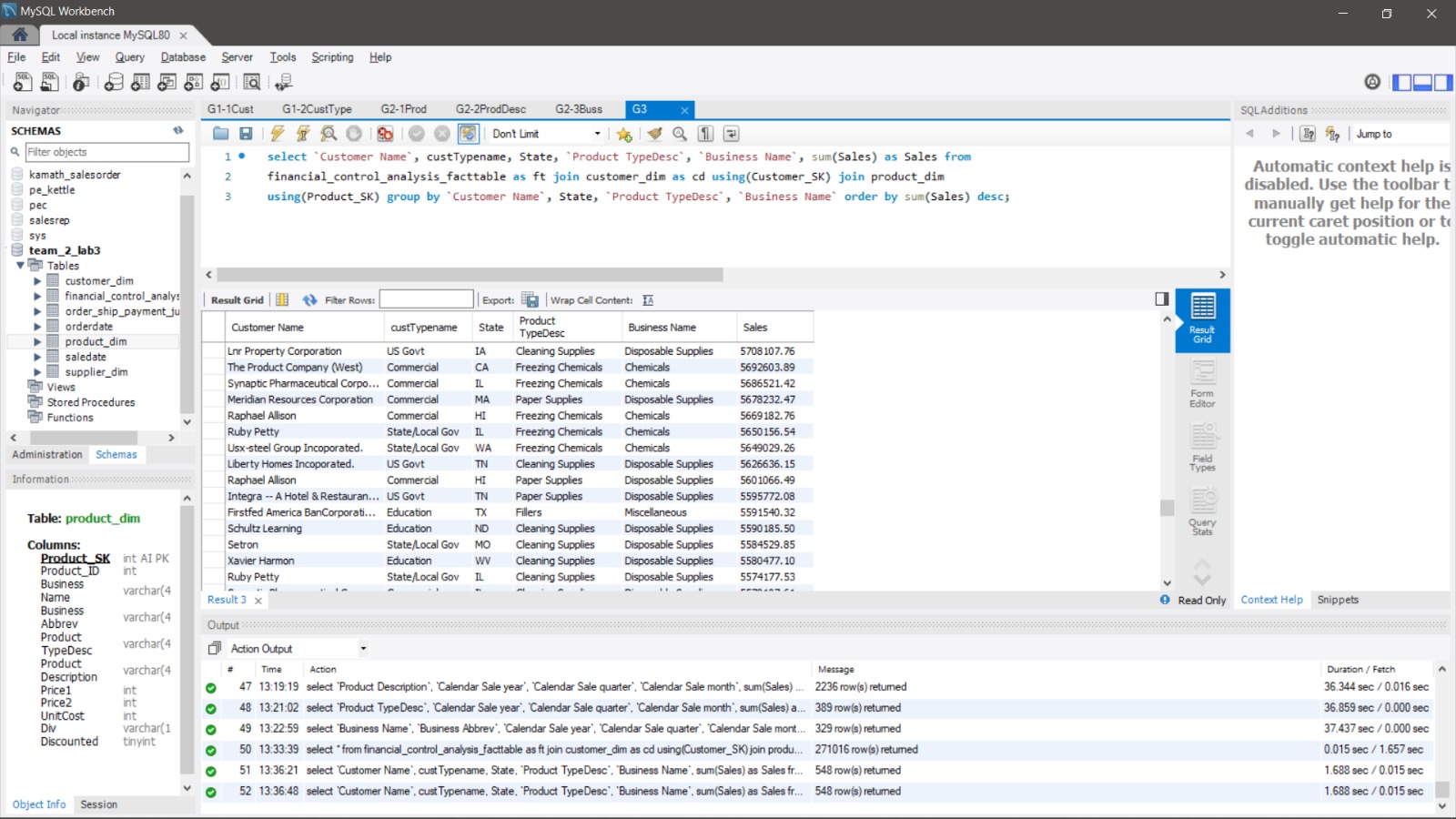
**DEFINING USER QUERIES:**

Based on lab2 queries, data provided in lab3 and the overall requirements, the following are the 3 user queries we have come up with:

**Goal 1: Sales by type of Customer, by state, by product type, and by business unit.**

**Query:**

SELECT `Customer Name`, custTypename, State, `Product TypeDesc`, `Business Name`, SUM(Sales) AS Sales FROM financial\_control\_analysis\_facttable AS ft JOIN customer\_dim AS cd USING(Customer\_SK) JOIN product\_dim USING(Product\_SK) GROUP BY `Customer Name`, State, `Product TypeDesc`, `Business Name` ORDER BY SUM(Sales) DESC;

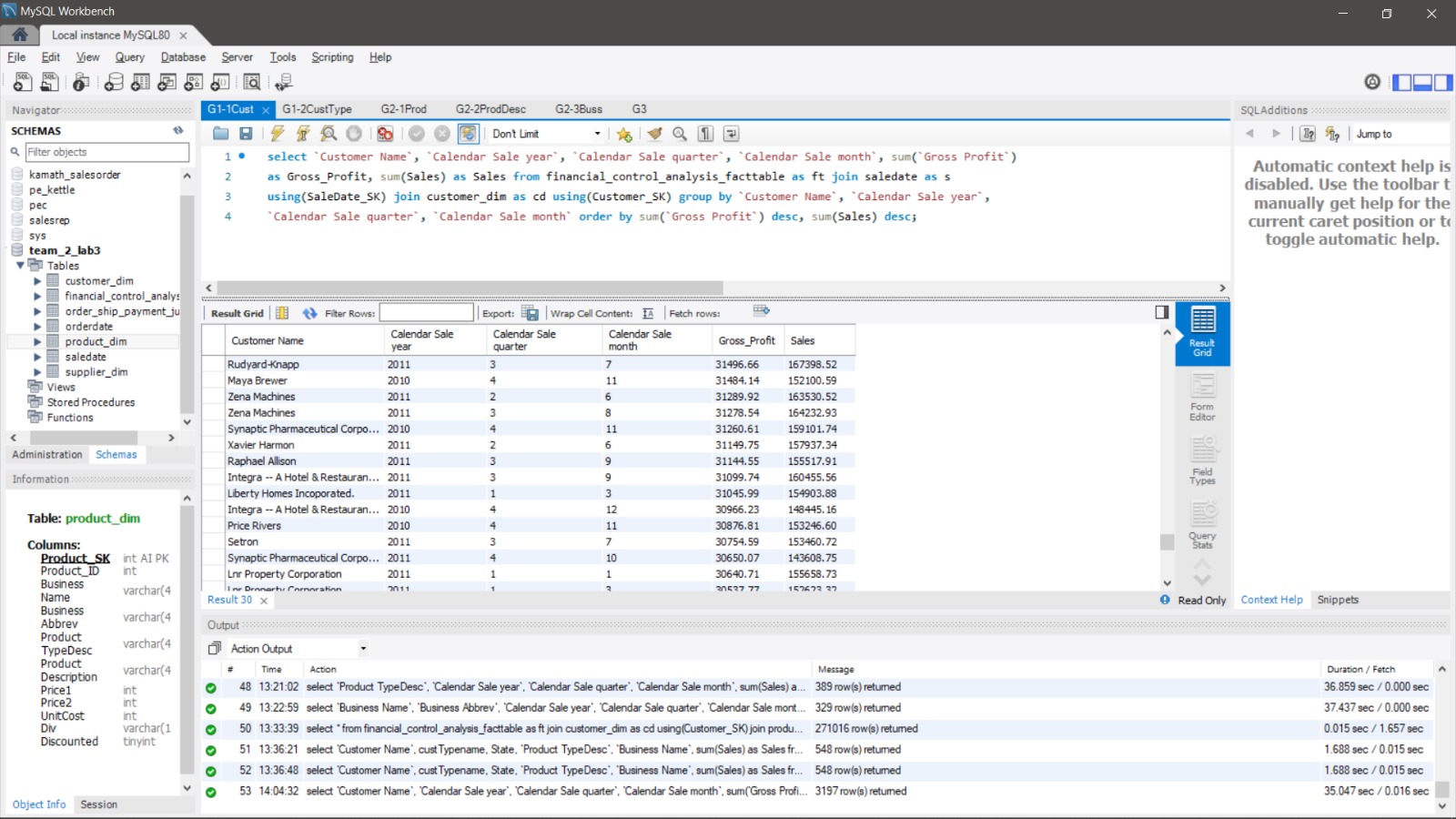


**GOAL 2: Show the sales, and the profits associated with each customer or customer type on an annual, quarterly, monthly basis.**

**First, we show the report annually, quarterly and monthly with each customer name.**

**Query:**

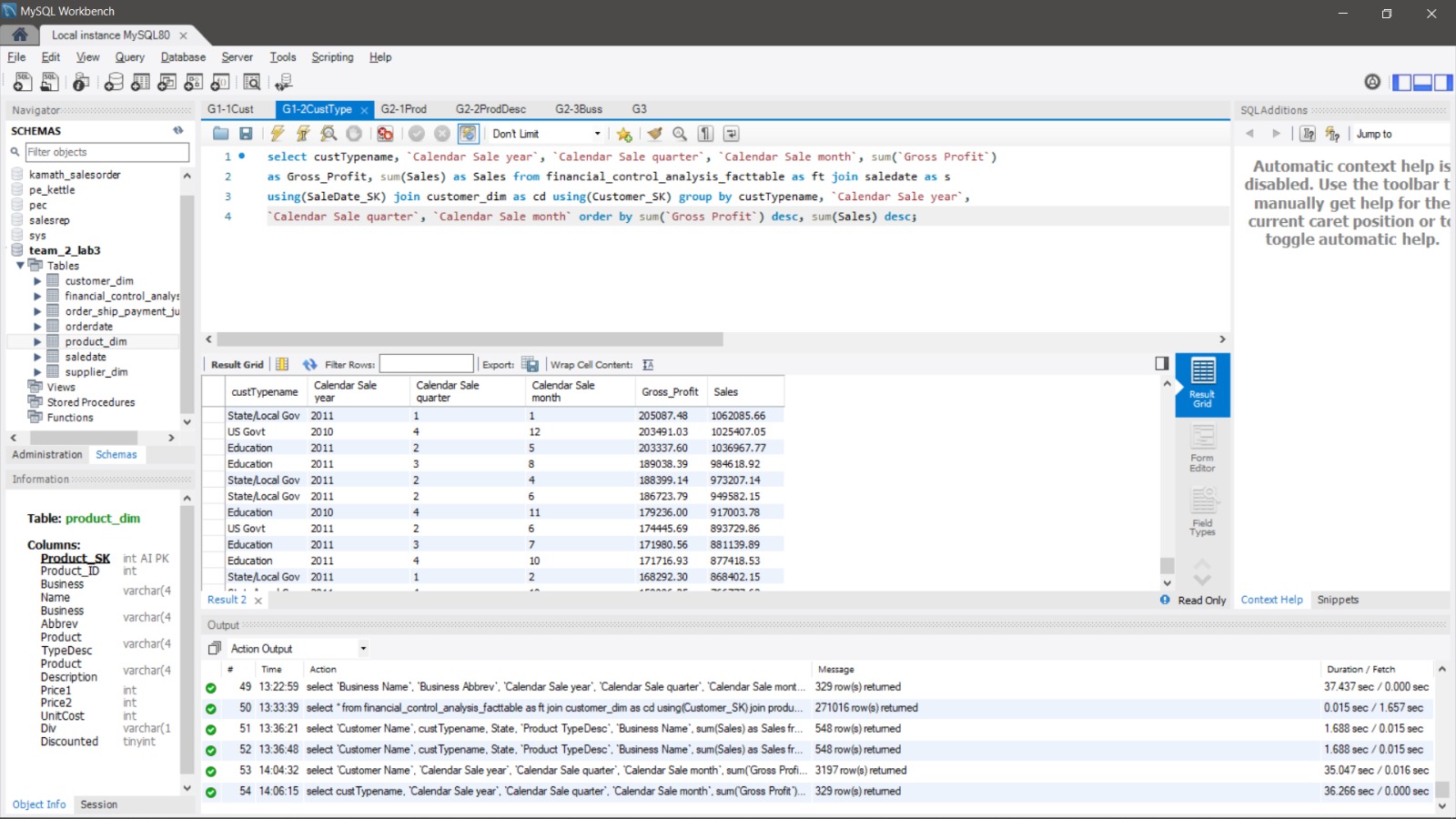
SELECT `Customer Name`, `Calendar Sale year`, `Calendar Sale quarter`, `Calendar Sale month`, SUM(`Gross Profit`) AS Gross\_Profit, SUM(Sales) AS Sales FROM financial\_control\_analysis\_facttable AS ft JOIN saledate AS s USING(SaleDate\_SK) JOIN customer\_dim AS cd USING(Customer\_SK) GROUP BY `Customer Name`, `Calendar Sale year`, `Calendar Sale quarter`, `Calendar Sale quarter` ORDER BY SUM(`Gross Profit`) DESC, sum(Sales) desc;



**Next, we calculate sales and profit per customer type.**

**Query:**

SELECT custTypename, `Calendar Sale year`, `Calendar Sale quarter`, `Calendar Sale month`, SUM(`Gross Profit`) AS Gross\_Profit, SUM(Sales) AS Sales FROM financial\_control\_analysis\_facttable AS ft JOIN saledate AS s USING(SaleDate\_SK) JOIN customer\_dim AS cd USING(Customer\_SK) GROUP BY custTypename, `Calendar Sale year`, `Calendar Sale quarter`, `Calendar Sale quarter` ORDER BY SUM(`Gross Profit`) DESC, SUM(Sales) desc;

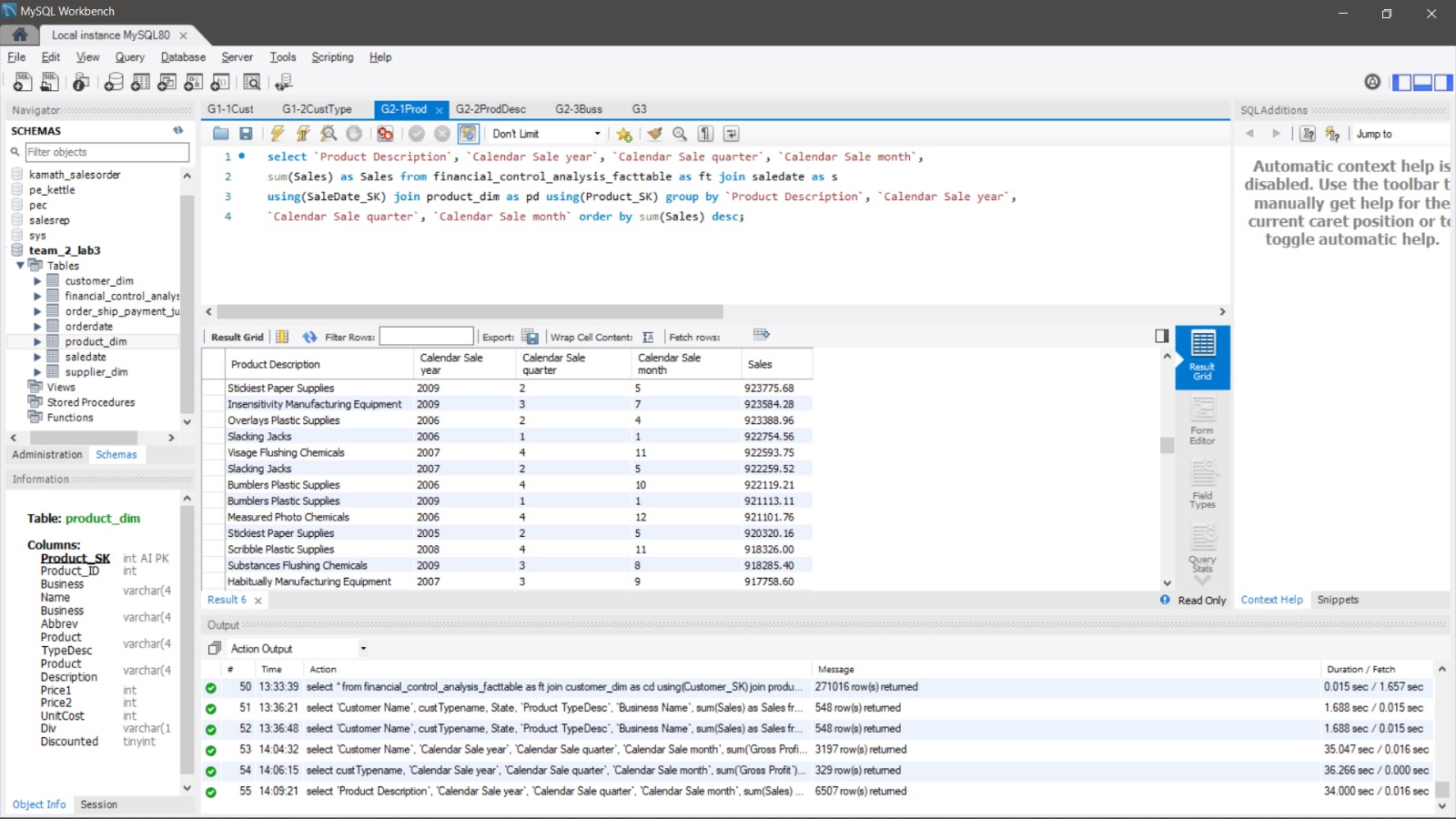


**Goal 3: Similar report as above but now on product description, product type and Business Unit**

**First, we calculate annual, quarterly and monthly sales fr product Description**

**Query:**

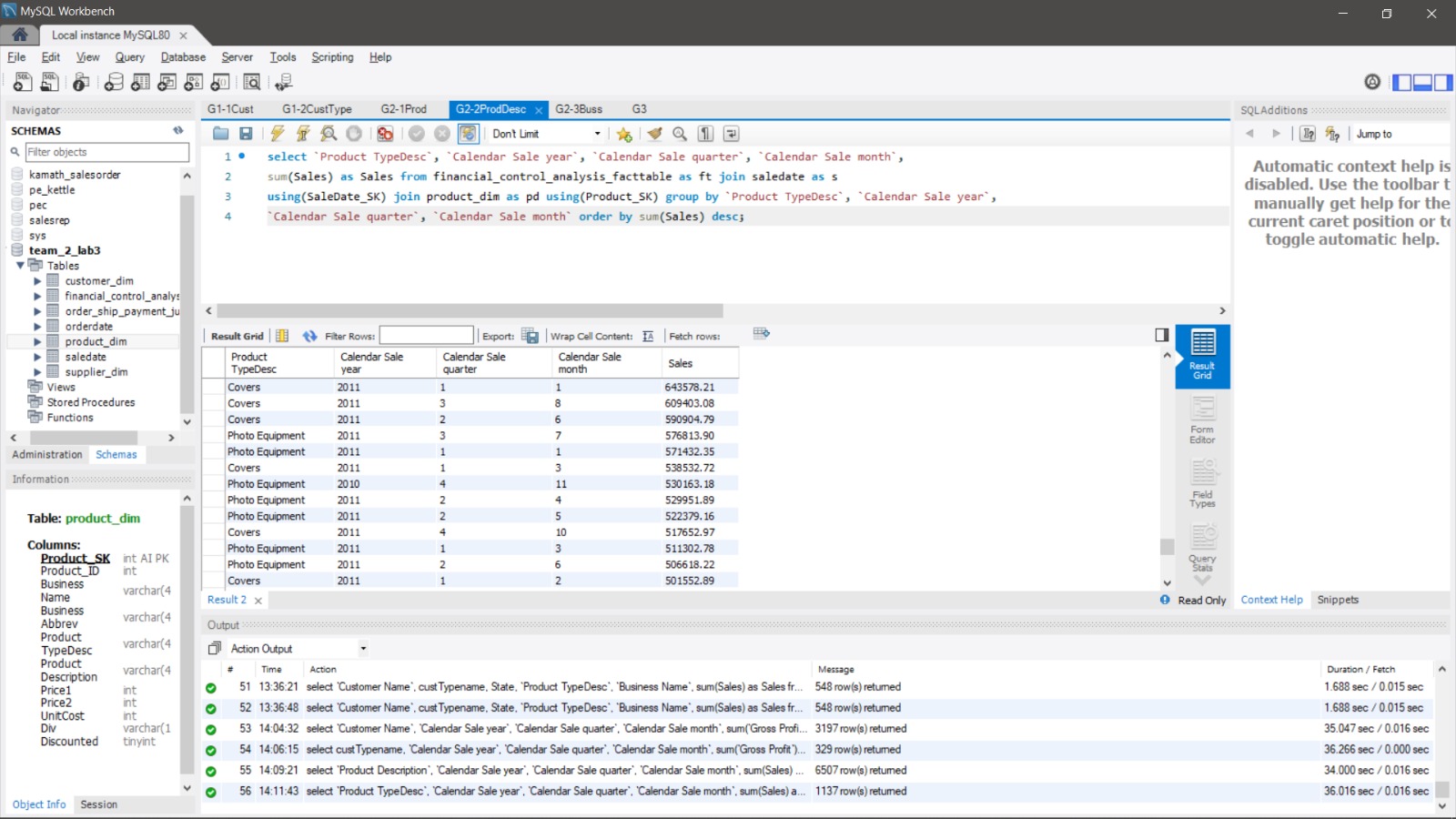
SELECT `Product Description`, `Calendar Sale year`, `Calendar Sale quarter`, `Calendar Sale month`, SUM(Sales) AS Sales FROM financial\_control\_analysis\_facttable AS ft JOIN saledate AS s USING(SaleDate\_SK) JOIN product\_dim AS pd USING(Product\_SK) GROUP BY `Product Description`, `Calendar Sale year`, `Calendar Sale quarter`, `Calendar Sale month` ORDER BY SUM(Sales) DESC;



**Next, we calculate the sales as above for Product Type Description:**

**Query:**

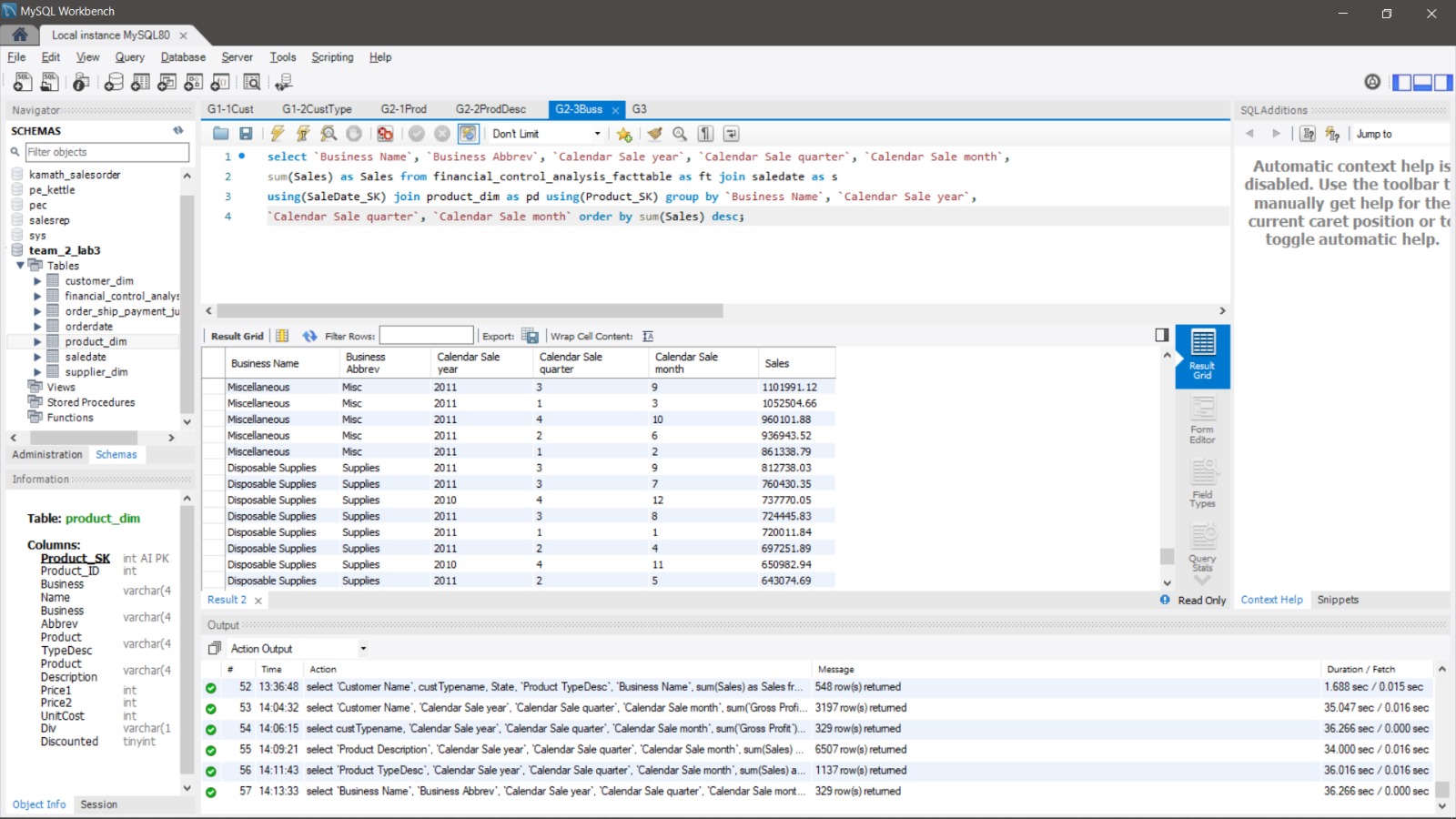
**SELECT `Product TypeDesc`, `Calendar Sale year`, `Calendar Sale quarter`, `Calendar Sale month`, SUM(Sales) AS Sales FROM financial\_control\_analysis\_facttable AS ft JOIN saledate AS s USING(SaleDate\_SK) JOIN product\_dim as pd USING(Product\_SK) GROUP BY `Product TypeDesc`, `Calendar Sale year`, `Calendar Sale quarter`, `Calendar Sale month` ORDER BY sum(Sales) DESC;**

****

**Finally, we calculate as above for business Unit:**

**Query:**

**SELECT `Business Name`, `Business Abbrev`, `Calendar Sale year`, `Calendar Sale quarter`, `Calendar Sale month`, SUM(Sales) AS Sales FROM financial\_control\_analysis\_facttable AS ft JOIN saledate AS s USING(SaleDate\_SK) JOIN product\_dim AS pd using(Product\_SK) GROUP BY `Business Name`, `Calendar Sale year`, `Calendar Sale quarter`, `Calendar Sale month` ORDER BY SUM(Sales) DESC;**

****