**Name \_\_\_Christopher Holmes\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Objective:**

* design a warehouse database using star schema design
* load a data warehouse table using sql against the production table
* creating a pivot table report in EXCEL
* using VISUAL EXPLAIN and QUERY STATISTICS to understand the steps involved in processing a select

**Additional References**

In asnwering the questions in part 6 you may want to read the additional references posted to this week’s module on iLearn about business intelligence and OLAP, and Big Data and NOSQL.

**Part 1: Create the HSD tables and load data**

Download files

* **Lab12 HSD CreateTablesWithData.sql**, the production tables for the HSD (Heather Sweeney Dodge Company) and
* **Lab12 HSDDW CreateTablesWithData.sql,**the data warehouse tables.

Execute the Lab12 HSD CreatetablesWithData script.

**Part 2: Create the HSDDW (HSD Data Warehouse) tables**

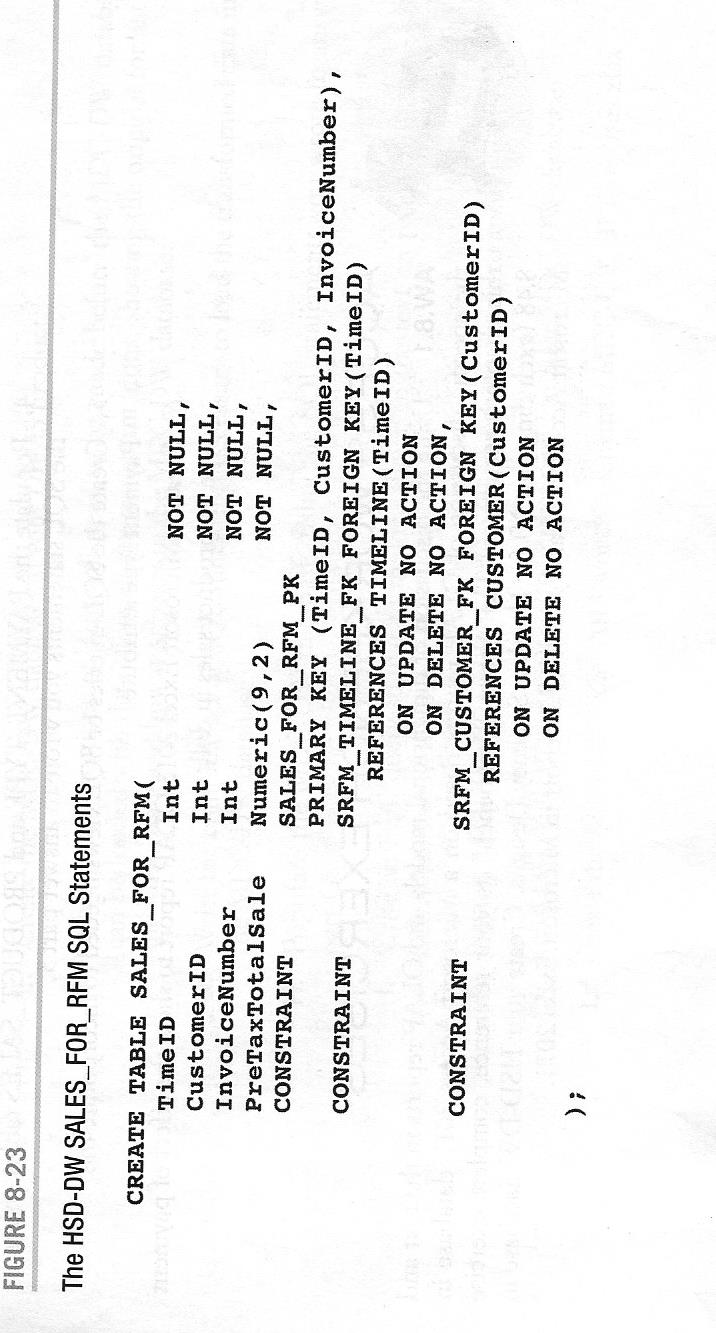
Execute the script Lab 12 HSDDW CreateTablesWithData which createx the warehouse tables and will loads the HSDDW tables with data from the HSD tables. Study the script carefully to understand how the data from HSD tables is copied and transformed and loaded into the HSDDW tables. This is known as ETL (Extract – Transform -Load). ETL may be done by sql scripts in simple cases (such as here). In more complex cases it needs special application programs to filter and “scrub” the data.

**Part 3: Modifying the HSD\_DW warehouse tables**

1. What transformations of data were made before HSD-DW was loaded with data? [List all the transformations, showing the original format of the HSD data and how they appear in the HSD-DW database.

|  |
| --- |
| For the customer table in the DW, the customer name was concatenated between the first name and the last name from the HSD; the domain of the email address was parsed from the email address; the area code was parsed from the full phone number. For the Product Sales table, the amount of items that the person purchased on the same day are added to a single entry with the total of all the sales. In the timeline table, the data is merged together to have the month, quarter id, quarter string, and the year. |

1. Create the SALEs\_FOR\_RFM table to the HSD-DW database using the create table statement below.



1. What transformations of data are necessary to load the SALES\_FORM\_RFM table?
2. What data will be used to load the SALES\_FOR\_RFM fact able? Write the complete set of SQL statements necessary to load the data.
3. Populate the SALES\_FOR\_RFM fact table using the SQL statement you wrote in the last question.
4. A query to summarize product units sold by Customer (CustomerName) , City, and Product(ProductName) and Year would be :

SELECT c.CustomerId, c.CustomerName, c.City,

p.ProductNumber, p.ProductName,

t.Year,t.QuarterText,

SUM(ps.Quantity) AS TotalQuantity

FROM customer c, product\_sales ps, product p,

timeline t

WHERE c.CustomerId = ps.CustomerID

AND p.ProductNumber = ps.ProductNumber

AND t.TimeId = ps.TimeID

GROUP BY c.CustomerId, c.CustomerName, c.City,

p.ProductNumber, p.ProductName,

t.QuarterText, t.Year

ORDER BY c.CustomerName, t.Year, t.QuarterText;

In the above query, product\_sales table is called the ***fact table***; customer, timeline and product tables are called ***dimension tables***.

Modify the above query to report the total dollar amount of each product for each year instead of the quantity sold for each product.

|  |
| --- |
| select c.CustomerId, c.CustomerName, c.City, p.ProductNumber, p.ProductName,  t.Year, t.QuarterText, SUM(ps.Quantity)\*(ps.unitprice) AS TotalDollarAmount  from customer c, product\_sales ps, product p, timeline t  where c.CustomerID = ps.CustomerID and p.ProductNumber = ps.ProductNumber and t.TimeId - ps.TimeID  group by c.CustomerId, c.CustomerName, c.City, p.ProductNumber, p.ProductName, t.QuarterText, t.Year, ps.unitprice  order by c.CustomerName, t.Year, t.QuarterText |

1. Write an SQL View equivalent of the SQL query you wrote in question 6

|  |
| --- |
| create view total\_per\_year as  select c.CustomerId, c.CustomerName, c.City, p.ProductNumber, p.ProductName,  t.Year, t.QuarterText, SUM(ps.Quantity)\*(ps.unitprice) AS TotalDollarAmount  from customer c, product\_sales ps, product p, timeline t  where c.CustomerID = ps.CustomerID and p.ProductNumber = ps.ProductNumber and t.TimeId - ps.TimeID  group by c.CustomerId, c.CustomerName, c.City, p.ProductNumber, p.ProductName, t.QuarterText, t.Year, ps.unitprice  order by c.CustomerName, t.Year, t.QuarterText |

1. Create the view that you defined in the last question.
2. Modify the design of the HSD-DW dimensional database to include a PAYMENT\_TYPE dimension table.
3. What data will be used to load the PAYMENT\_TYPE table? Write the complete set of SQL necessary to load these data.

**Part 4: Create a pivot table in EXCEL**

***[Note: your version of EXCEL may be slightly different from what is described in the instructions below]***

You must be running a Windows computer to do part 4. You need to have ODBC connector installed and configured. Watch the video on insalling and configuring ODBC.

If you do not have a Windows computer, skip to Part 4a.

Using the DATA tab in EXCEL, select “Get External Data” and then “Microsoft query” and then select the odbc data source for the hsddw database.

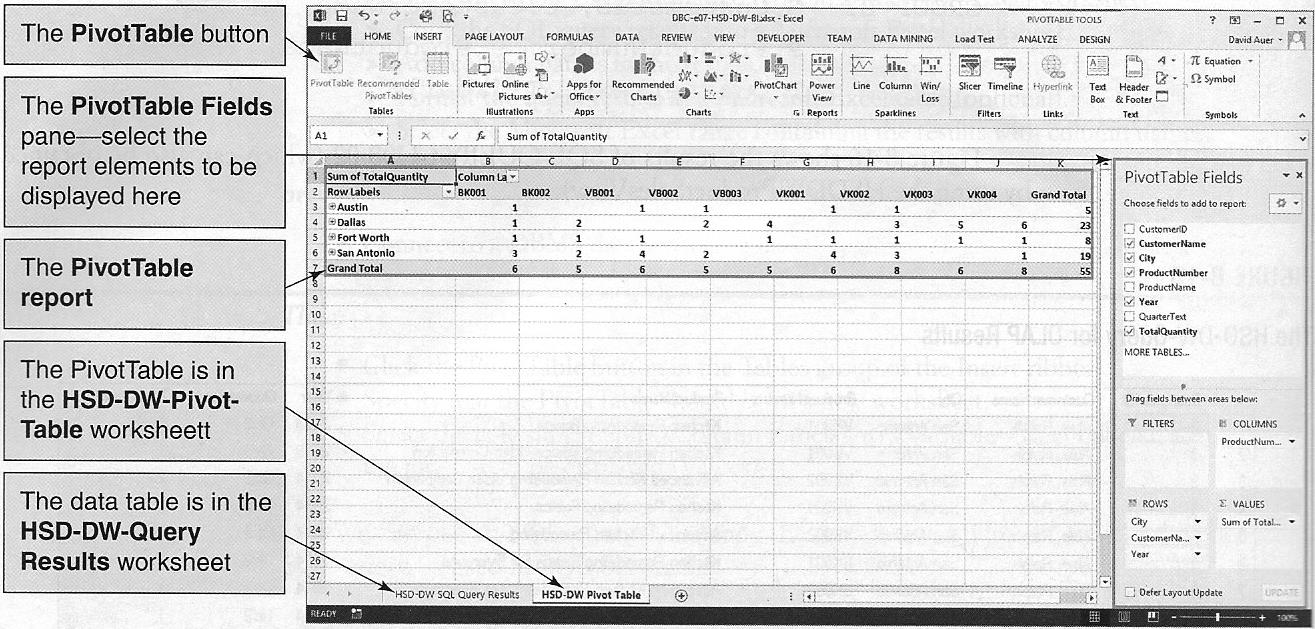
Select the table and columns for the query in Part 3 #6 that summarizes by TotalQuantity. You can select the tables, column and specify the ordering using the Microsoft Query dialogs but you will have to finish editing the query. The final query should look like the query in problem 6 in part 3.

Execute the query and then return data to the EXCEL Worksheet as TABLE.

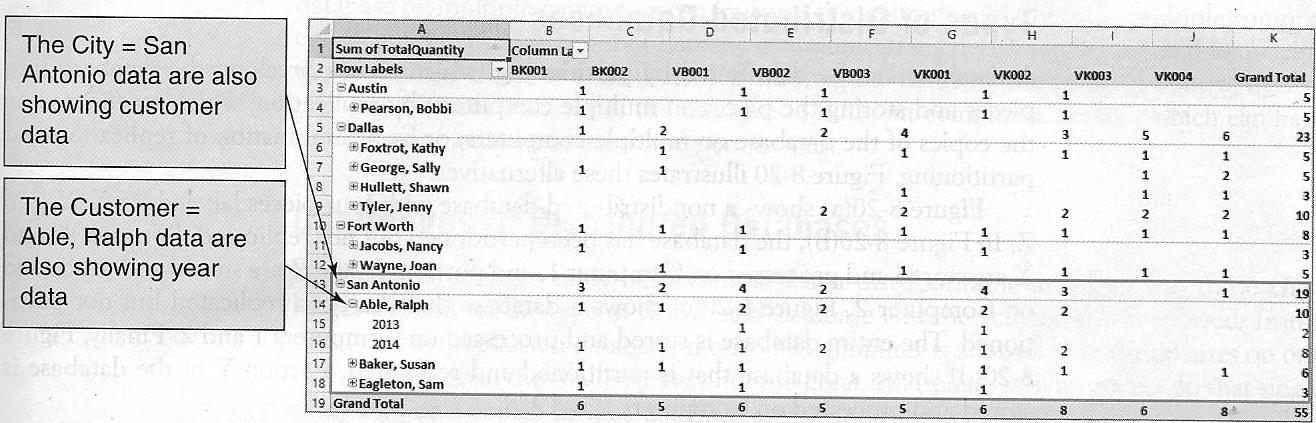
Go the INSERT tab and use the PIVOT table tool.

Drag fields City, CustomerName and Year into the ROWS panel and ProductNumber into the COLUMNS panel and TOTALQUANTITY field into the VALUES panel. Your pivot table should look like the following.

OLAP ProductNumber by City Report



You can expand (called drilling down on) each City to get details on CustomerName and Year.



Take a screenshot of your pivot table and paste here.

**Part 4a: Create a pivot table in Google Sheets**

Using MySql Workbench perform the query in Part 3 #6 that summarizes by customer, city, quarter and total dollar amount sold. Export the result into a csv file.

Create a blank google sheet and do File 🡪 Import. Select the upload option and upload the csv file.

Next do Data 🡪 Pivot Table.

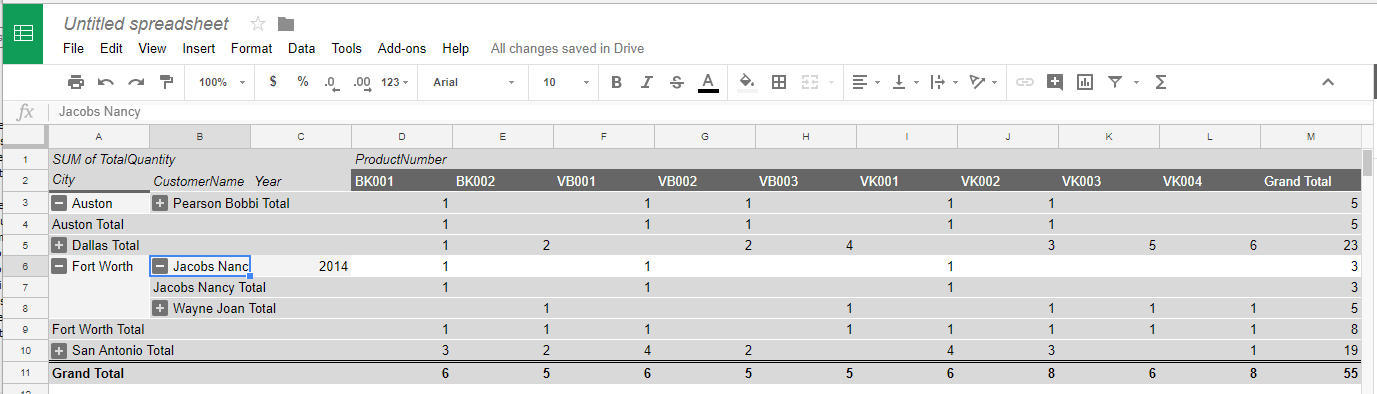
Click Add Rows and add City, CustomerName and Year.

Click Add Columns and add ProductNumber.

Finally Click Add Values and add TOTALQUANTITY.

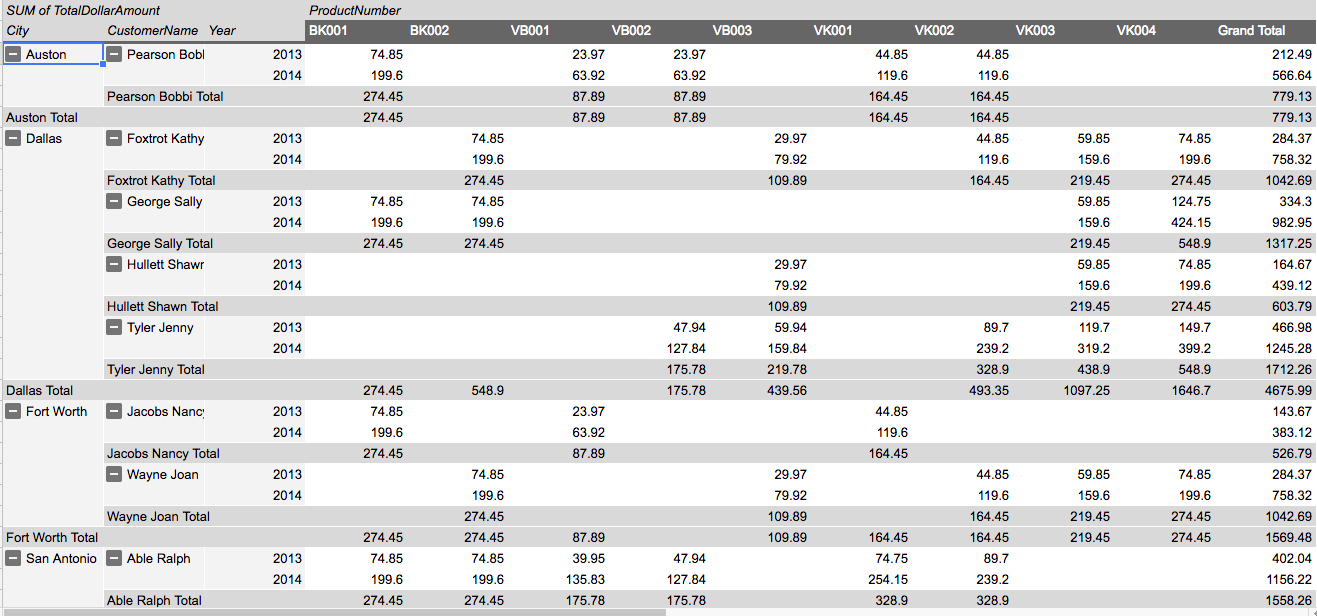
Your pivot table should look like the following.

OLAP ProductNumber by City Report



You can expand (called drilling down on) or collapse each City to get details on CustomerName, Year and Product.

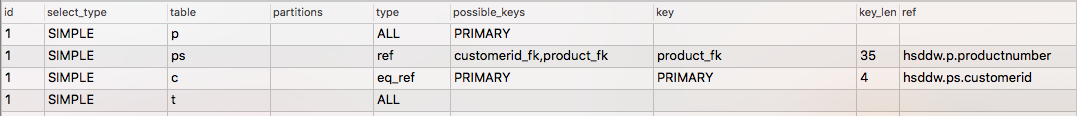
Take a screenshot of your pivot table and paste here.

****

**Part 5: Visual Explain and Query statistics**

Copy and paste the query from part 3 question 6 into a query tab in MySQL Workbench. Execute the query scroll down the right-hand side of the result panel and click “Execution Plan” and “Query Stats”.

Paste a screenshot of the execution plan here.



Explain in words what the execution plan shows and the query stats. [Consult the MySQL online reference or ask in the Q&A forum if you need help reading the execution plan]

For your information:

*When a query is taking a very long time, one of the responsibilities of a DBA often will be to investigate how the server is processing the query and why it is taking so long and what can be done to decrease the time. The visual explain and query stats are used to analyze how the server is processing the select statement and may lead some ideas to improve performance. Some things that a DBA might do to improve performance include:*

* *reorganize the table using OPTIMIZE TABLE command so the table rows are in physical sequence on the HDD by primary key*
* *Make use of VARCHAR datatype (instead of CHAR), use COMPRESSED or COMPACT options to reduce size of rows and tables to reduce the number of IOs and time needed to scan the table*
* *Create additional indexes or rearrange the column order of a multi-column index.*
* *How a select statement is processed (which indexes are used, how multiple indexes are used, how much of the table has to scanned and filtered, is the table and index files accessed in sequential or random order) depends on*
* *the sql statement*
* *the table statistics*
* *DBMS server type and version*
* *DBMS Configuration options*

**Part 6 : Additional questions.**

1. What is a BI (Business Intelligence) system?

|  |
| --- |
| A Business Intelligence system is a system that utilizes a technology-driven process to be able to analyze data and create actionable information that decision makers in a business can use to make informed decisions. |

1. How does a BI system differ from a transaction processing system?

|  |
| --- |
| A BI System does not support the same operational activities that a transaction processing system would; such as recording and processing orders |

1. What is an ETL system, and what functions does it perform?

|  |
| --- |
| An ETL system is a system that will read data from an operational database. |

1. What is the enterprise data warehouse (EDW) architecture?

|  |
| --- |
| Is a combination of the data mart structure and data warehouse architecture. The data warehouse is the authority and maintains all enterprise BI data. The data mart gets all the data from the data warehouse and does not add any data. |

1. What is a star schema?

|  |
| --- |
| A star schema is a design that will use a denormalized design that stores historical data. There is a fact table at the center, and all the other dimensional tables radiating out from that center. A fact table will always be fully normalized, but a dimension table may be not. |

1. What does OLAP stand for?

|  |
| --- |
| Online Analytical Processing |

1. What are the distinguishing features of a OLAP reports?

|  |
| --- |
| -Produce OLAP reports  -Dynamic allowing a user to change the format |

1. What is *Big Data*?

|  |
| --- |
| Big data is the term that is used for extremely large datasets that can be generated by web applications; Facebook, Google, Twitter. |

1. What is the *NOSQL* movement?

|  |
| --- |
| The NOSQL movement is the movement to have the big data stored in nonrelational dmb’s. |