Data Warehousing using SLQ

Questions

1. You have been asked to integrate two sets of data from two sales agents of the same company into a single source that can be analyzed. The two extracted sets of data can be found in the source 1 and source 2 worksheet in Excel. It is a collection of orders, organized by item in each order. Multiple rows can be associated with a single order because you can order multiple items in a single order. The column "extended price" is the actual price paid by a customer. The company distinguishes each customer into three different customer statuses: "Silver," "Gold," and "Platinum." Platinum is the best.

To understand the inconsistencies between the data, open the workbook and look at the Source 1 and Source 2 worksheets. You'll notice that the data doesn't quite match up. Please identify the inconsistencies of the two data sources and propose what you would do before loading the data into the company's data warehouse. Please structure your answers according to the following format.

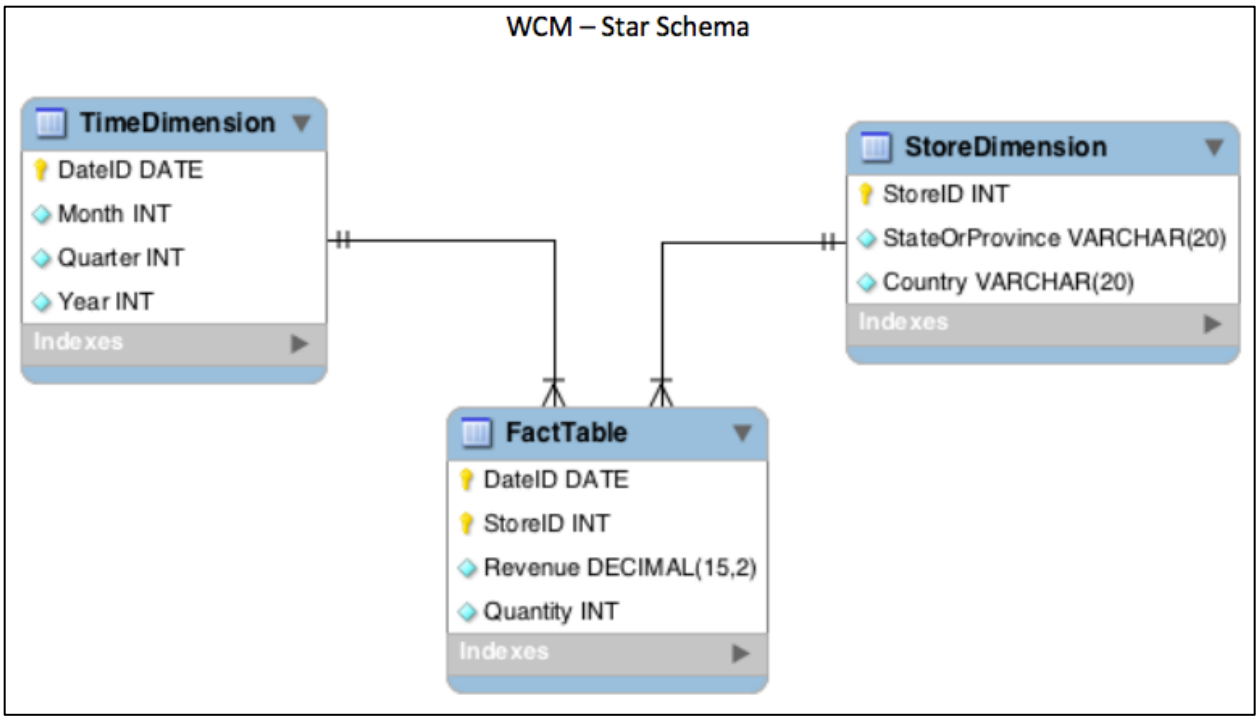
Source 1 Field	Source 2 Field	Solution
Column Names • Customer Name • Order Date • Product • Product ID • Unit Price • Extended Price • Full Price	Column Names • Customer First Name • Customer Last Name • Order Date • Product • Product ID • Unit Price • Extended Price • Full Price	Columns need to match from both sources in terms of names and formatting
Order ID: 5-digit number	Order ID: A + 5 digits	Standard format
Customer State/Province: States are abbreviated	Customer State/Province: States are fully spelled out	Standard format
Customer Status: Ranked in Silver, Gold, and Platinum being the best	Customer Status: Ranked in number- 1,2,3	Make sure both sources are ranked in silver, gold, and platinum only. Ranking in

		Numbers are very vague. 3 could be the best or worst.
Unit Price, Full Price, Extended Price, Total Discount,	Unit Price, Full Price, Extended Price	Source 1 has Total Discount column while Source 2 doesn't. Both sources should have it

2. Use the attached data file: WCM Movie Rentals.csv. You can Load the file (the World Classic Movie Rentals history) in Excel and examine the column names. The column names are self-explanatory. For each of the columns listed below, identify the column as a measure or a dimension:

	Column	**Measure or Dimension?**
1	Year	dimension
2	Quarter	dimension
3	Month	dimension
4	Country	dimension
5	StateOrProvince	dimension
6	Revenue	Measure
7	Quantity	Measure

3. The data in the WCM Movie Rentals.csv file is generated from a database that has the star schema as shown below. Write an SQL query that can generate this data from the database. *Note that in the fact table of the star schema, each row is at the day-store level; in the csv file, each row is at the month-state level.*




```
SELECT
t.year AS year,
t.month AS month,
s.state AS state,
SUM(f.revenue) AS total_revenue
FROM
fact_table f
JOIN
fact_table f ON f.store_id=s.store_id
JOIN
time_dim t ON f.date_id=t.date_id
GROUP BY
t.year,t.month,t.state
ORDER BY
t.year,t.month,s.state;
```

OLAP Queries

4. Create a data cube that consists of revenues generated from each city, country combination; order the results in descending order of revenues. Based on your result, which city produced the highest revenues?

Query:

```sql
SELECT
    cd.Country,
    cd.City,
    SUM(ft.Revenue) AS total_revenue
FROM
    FactTable ft, CustomerDimension cd
WHERE
ft.CustomerID = cd.CustomerID
GROUP BY
    cd.Country, cd.City
ORDER BY
    total_revenue DESC;
```

Result: Based on your results, which city produced the highest revenues?

Cunewalde

5. You are interested in understanding the sales trend over time. Please write a query to generate a data cube that would list total revenues in each month over time.

```sql
SELECT
    td.Year,
    td.Month,
    SUM(ft.Revenue) AS total_revenue
FROM
    FactTable ft, TimeDimension td
WHERE
ft.DateID = td.DateID
GROUP BY
    td.Year, td.Month
ORDER BY
    total_revenue DESC;
```

6. Generate a data cube that shows product category sales in each quarter of 1997. In order to see if there is clearly sales trend for each product category, you want to show category first and then quarter, and then the quarterly sales of the category.

```sql
SELECT
    pd.CategoryName,
    td.Quarter,
    SUM(ft.Revenue) AS Quartely_Sales
FROM
    FactTable ft
JOIN
    ProductDimension pd ON ft.ProductID = pd.ProductID
JOIN
    TimeDimension td ON ft.DateID = td.DateID
WHERE
    td.year = 1997
GROUP BY
    pd.CategoryName, td.Quarter
ORDER BY
    pd.CategoryName, td.Quarter;
```