Data Warehouse Design Background

This document provides details about the requirements of the data warehouse for CPI Card Group[1]. The requirements provide details about the design and sizes of data sources as well as business reporting needs.

Data Sources

The data warehouse uses three data sources as depicted in Figure 1. The ERP database is the major data source used by manufacturing to manage jobs, subjobs, shipments, and invoices. The lead file and financial summaries are secondary data sources, both in spreadsheet format. The lead file and financial summary are prepared from other data sources used by the marketing and accounting departments.

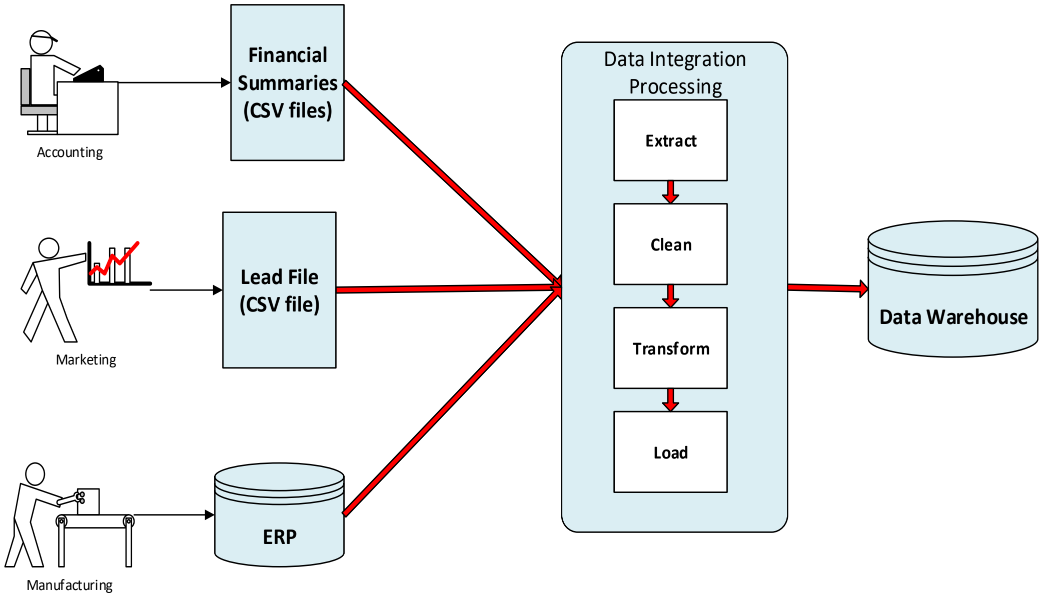

Figure 1: Data Sources for the ABC Data Warehouse

[1] The details are based on the requirements faced by CPI Card Group but have been simplified for usage in this course.

ERP Database Design

The ERP database supports complete processing for jobs involving planning, manufacturing, shipping, invoicing, and payment processing as well as accounting. However, the complete details are not important for this case. Figure 2 shows an abbreviated ERD for the subset of the ERP database relevant for the initial phase of the data warehouse. Table 1 provides a brief description of each table. Appendices A and B contain a complete ERD and details about each column.

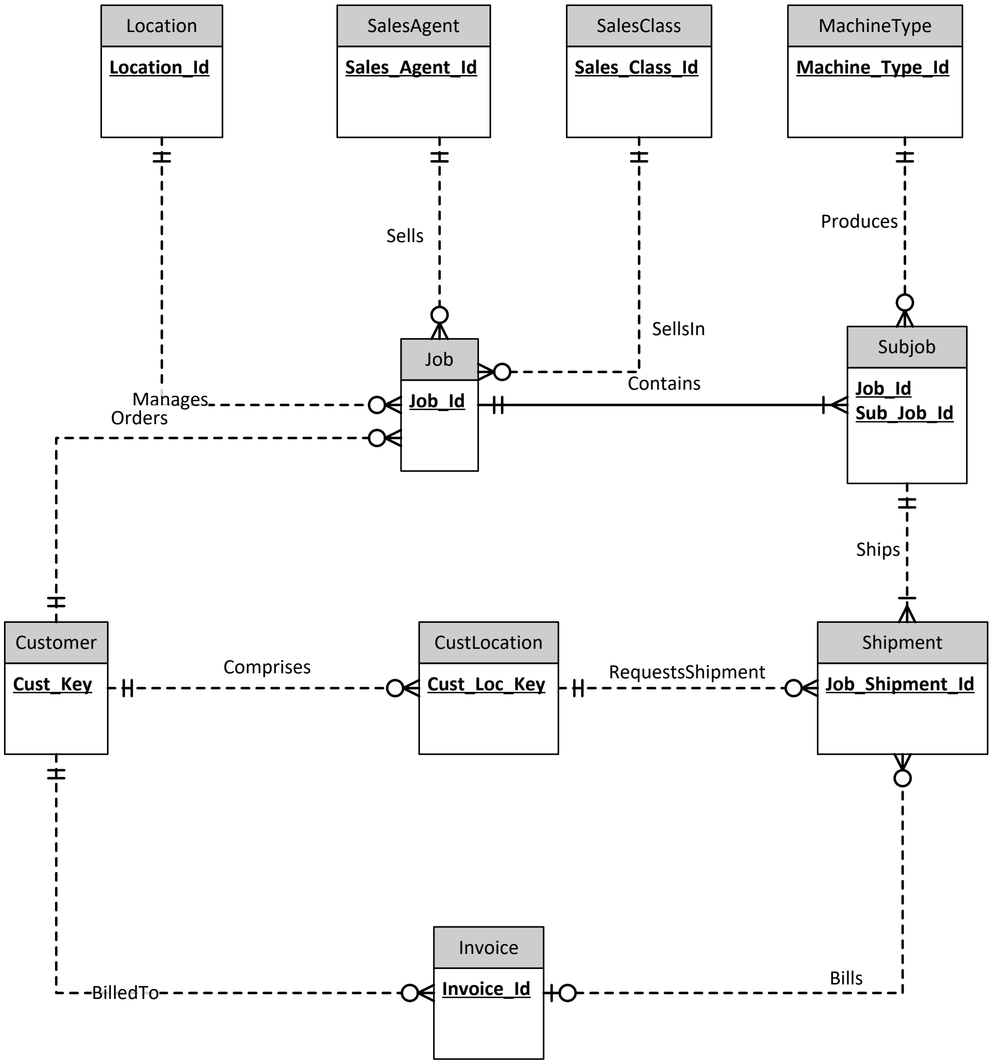

Figure 2: Abbreviated ERD for a Subset of the ERP Database

Table 1: Entity Type Definitions

Entity Type	Comments
Customer	Organizations that request jobs. Customers are involved in quotes which are recorded in the CRM not in the ERP. In the ERP, customers are recorded in a job.
CustLocation	Locations of customers to which cards are shipped.
Invoice	Collection of shipments billed to a customer. An invoice is created after related shipments so the Bills relationship is optional.
Job	A contract for a quantity of cards generated after a customer accepts a quote
Location	Location of the company that manages a job
MachineType	Type of machine used to produce cards in a subjob
SalesAgent	Employee credited with obtaining a job
SalesClass	Type of product on a job
Shipment	Collection of cards shipped to a customer after production in a subjob
SubJob	Subset of a job produced using a machine type. Identification dependent on Job.

The secondary data sources are simpler than the ERP database. Appendices C to E contain details about the secondary data sources.

Sample Data

To clarify the data sources, sample data are provided for the tables of the ERP database as well as the other data sources. Due to number of columns and long column names, the sample data are contained in a separate spreadsheet.

Data Source Size Statistics

To compute grain size, you should use the estimates about cardinalities of tables and unique values of some columns shown in Table 2. The number of rows reflects the current sizes of the data sources with two years of data. For the job, subjob, shipment, and invoice tables, the sizes reflect one year of data as the rows in these tables are archived after one year.

Table 2: Size Estimates for Data Sources

Table	Rows	Comments
Customer	3,000	300 postal codes
CustLocation	10,000	20% of customers have 1 location; remainder have about 4 locations each; 500 postal codes
Invoice	1,000,000	2.5 shipments per invoice
Job	100,000	40% of leads turn into jobs, 50,000 jobs per year
Location	10	
MachineType	10	
SalesAgent	50	
SalesClass	6	
Shipment	2,500,000	5 shipments per subjob on average
SubJob	500,000	5 subjobs per job on average
Lead file	250,000	About 125,000 leads per year
Financial Sales Summary	1,800	Monthly summary for combinations of sales class and location for 5 years. Assumes sparsity of 70%
Financial Cost Summary	5,400	Monthly summary for combinations of sales class, location, and machine type for 5 years. Assumes sparsity of 85%

Business Reporting Needs

The main purpose of the data warehouse is to track and compare sales and costs for major dimensions across time periods. Sales should also be compared to invoiced amounts for major dimensions and time periods. Costs should be tracked by component for labor, machine, overhead, and material so that standard accounting measures can be computed such as gross margin, contribution ratio, and related ratios. In addition, planning performance should be evaluated by comparing sales to forecasts and costs to budgets.

Job and Shipment Performance and Trends

- What are job revenue trends by location over time?
- What are sales agent productivity from leads to jobs over time?
- What are production trends for jobs (time to subjob production) for entities over time?
- What are shipment trends for jobs (contract time to shipment) for entities over time as compared to shipment promised dates and first shipping dates?

Invoice Trends

- Which entities (such as customers, locations, and products) generate the highest invoice amounts over time?
- What are trends for invoicing of job amounts (time to invoice) for entities (locations and products) over time?
- What are trends over time for returns measured by the difference between invoice quantity and shipping quantity for products, machines, and locations?

Financial Performance

- What are the gross margins for a location?
- How much does a location’s gross margin vary from its forecast/budget by month?
- What products are the most difficult to budget or forecast?
- What products and locations are the most profitable over time?

Appendix A: Complete ERD for the ERP Database

The complete ERD (Figure 1) shows all columns including primary keys, foreign keys, and other columns in each table. Note that the primary key of Subjob is a combination of Job_Id and Sub_Job_Id. Foreign keys in bold font indicate that the column is required.

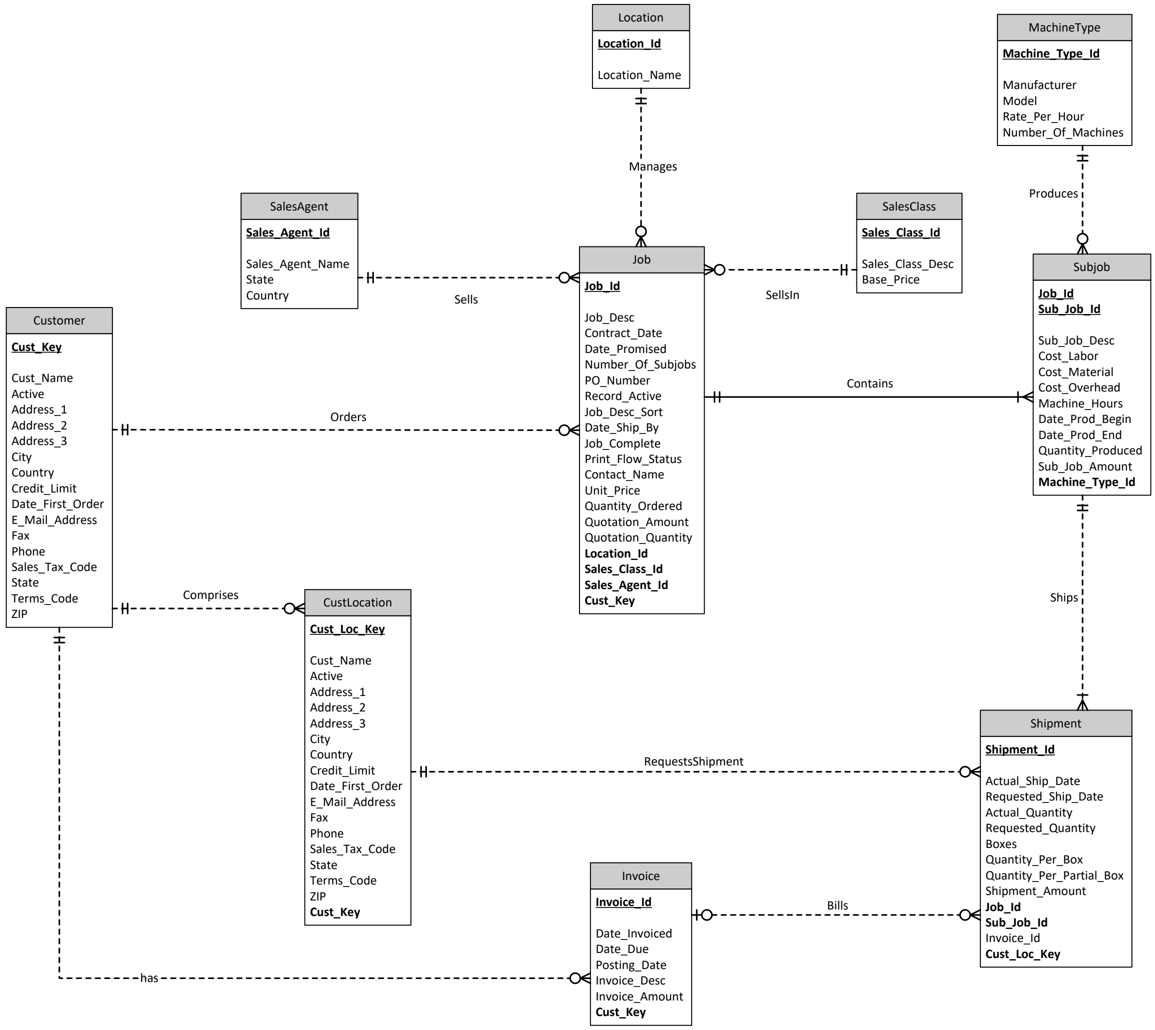

Figure A.1: Complete ERD for a Subset of the ERP Database

Appendix B: Data Dictionary for ERP Columns

The tables in this appendix contain selected details about each column in the ERP database. The DW column indicates if the column has likely value for the data warehouse. You should also see the spreadsheet with example values for each table.

Customer

Column Name	Data Type	Definition	DW
Cust_Key	LONG	unique identifier for customer	
Cust_Name	VARCHAR	customer name	
Active	BOOLEAN	marks whether the customer is active	No
Address_1	VARCHAR	customer address line 1	No
Address_2	VARCHAR	customer address line 2	No
Address_3	VARCHAR	customer address line 3	No
City	VARCHAR	customer city	
Country	VARCHAR	customer country	
Credit_Limit	CURRENCY	customer credit limit	
Date_First_Order	DATE	date of the customer's first order	No
E_Mail_Address	VARCHAR	customer e-mail address	
Fax	CHAR(10)	customer fax number	No
Phone	CHAR(10)	customer phone number	Maybe if parsed
Sales_Tax_Code	CHAR(10)	customer sales tax code	No
State	CHAR(2)	customer state	
Terms_Code	CHAR(10)	Indicates payment terms	
ZIP	CHAR(10)	customer ZIP code	

CustLocation

Column Name	Data Type	Definition	DW
Cust_Loc_Key	LONG	unique identifier for customer location	
Cust_Name	VARCHAR	customer location name	
Active	BOOLEAN	marks whether the customer location is active	No
Address_1	VARCHAR	customer location address line 1	No
Address_2	VARCHAR	customer location address line 2	No
Address_3	VARCHAR	customer location address line 3	No
City	VARCHAR	customer location city	
Country	VARCHAR	customer location country	
Credit_Limit	CURRENCY	customer location credit limit	
Date_First_Order	DATE	date of the customer location's first order	No

E_Mail_Address	VARCHAR	customer location e-mail address	
Fax	CHAR(10)	customer location fax number	No
Phone	CHAR(10)	customer location phone number	Maybe if parsed
Sales_Tax_Code	CHAR(10)	customer location sales tax code; only used for non-commercial customers	No
State	CHAR(2)	customer location state	
Terms_Code	CHAR(10)	customer payment terms	
ZIP	CHAR(10)	customer location ZIP code	
Cust_Key	LONG	identifier of the customer	

Invoice

Column Name	Data Type	Definition	DW
Invoice_Id	LONG	Unique identifier of the shipment	
Date_Invoiced	DATE	Date the invoice was prepared	
Date_Due	DATE	Date the payment should be received; depends on payment terms	
Posting_Date	DATE	Date the payment was recorded	No
Invoice_Desc	VARCHAR	Description of the invoice contents	No
Invoice_Amount	CURRENCY	Amount of invoice	
Invoice_Quantity	INTEGER	Quantity billed on invoice	
Invoice_Shipped	INTEGER	Quantity sent in related shipments; Returns are difference between invoice quantity and shipped	
Cust_Key	LONG	identifier of the customer billed on the invoice	

Job

Column Name	Data Type	Definition	DW
Job_Id	LONG	Unique identifier of the job	
Job_Desc	CHAR(50)	Description of the job. Used only when a job is not complete.	
Contract_Date	DATE	Date the job contract was created	
Date_Promised	DATE	Date promised for the last shipment of the the job	
Number_Of_Subjobs	INTEGER	Number of subjobs associated with the job. The number of subjobs is initially estimated but then updated if the number changes during production.	
PO_Number	CHAR(10)	Purchase order number of the job from the customer	No
Record_Active	BOOLEAN	True if the job is active. Only non-active jobs will be stored in the data warehouse.	No

Date_Ship_By	DATE	Date promised for the first shipment	
Job_Complete	BOOLEAN	True if job is complete. Usually the same value as Record_Active.	No
Print_Flow_Status	CHAR(10)	Indicates production status. Not useful after a job is completed.	No
Contact_Name	CHAR(50)	Name of the contact for the job.	No
Unit_Price	CURRENCY	Price of each unit created for the job	
Quantity_Ordered	SHORT	Number of items ordered	
Quotation_Amount	CURRENCY	Dollar amount of the quote to the customer	
Quotation_Ordered	SHORT	Number of items initially requested by the customer	
Location_Id	LONG	Identifier of the location where the job belongs	
Sales_Class_Id	LONG	Identifier of the sales class where the job belongs	
Sales_Agent_Id	LONG	Identifier of the sales agent associated with the job	
Cust_Key	LONG	Identifier of the customer who placed the order	

SubJob

Column Name	Data Type	Definition	DW
Job_Id	LONG	identifier of the job and part of the primary key	
Sub_Job_Id	SHORT	identifier of the subjob within the job	
Sub_Job_Desc	CHAR(50)	description of the subjob	No
Cost_Labor	CURRENCY	cost of labor for the subjob	
Cost_Material	CURRENCY	cost of materials for the subjob	
Cost_Overhead	CURRENCY	cost of overhead for the subjob	
Machine_Hours	DECIMAL	number of machine hours used for the subjob	
Date_Prod_Begin	DATE	date the production of the subjob began	
Date_Prod_End	DATE	date the production of the subjob ended	
Quantity_Produced	INTEGER	number of items produced for the subjob	
Sub_Job_Amount	CURRENCY	dollar value of the items produced for the subjob	
Machine_Type_Id	LONG	identifier of the machine type used for the subjob	

Shipment

Column Name	Data Type	Definition	DW
Shipment_Id	LONG	unique identifier of the shipment	
Actual_Ship_Date	DATE	date the shipment actually occurred	
Requested_Ship_Date	DATE	date the shipment was requested by the customer	

Actual_Quantity	INTEGER	actual quantity of items shipped	
Requested_Quantity	INTEGER	requested quantity of items shipped	
Boxes	INTEGER	number of full boxes in the shipment	
Quantity_Per_Box	INTEGER	number of items in each box	
Quantity_Per_Partial_Box	INTEGER	number of items in the partially filled box	
Job_Id	LONG	identifier of the job related to the shipment	
Shipment_Amount	CURRENCY	Amount to be billed for shipment	
Sub_Job_Id	LONG	identifier of the subjob related to the shipment	
Invoice_Id	LONG	identifier of the invoice related to the shipment; null until invoiced	
Cust_Loc_Key	LONG	identifier of the related customer location	

Location

Column Name	Data Type	Definition	DW
Location_Id	LONG	unique identifier for the location	
Location_Name	CHAR(50)	name of the location	

MachineType

Column Name	Data Type	Definition	DW
Machine_Type_Id	LONG	unique identifier for machine type	
Manufacturer	LONG	manufacturing company of the machine type	
Model	VARCHAR	specific model of the machine type	
Rate_Per_Hour	CURRENCY	Rate per hour charged for using machine type	
Number_Of_Machines	INTEGER	number of available machines	

SalesAgent

Column Name	Data Type	Definition	DW
Sales_Agent_Id	LONG	unique identifier for sales agent	
Sales_Agent_Name	VARCHAR	sales agent name	
State	CHAR(2)	sales agent state	
Country	CHAR(25)	sales agent country	
Record_Active	BOOLEAN	True if the sales agent is active with the company	No

SalesClass

Column	Data Type	Definition
Sales_Class_Id	LONG	unique identifier for sales class

Sales_Class_Desc	VARCHAR	type of card produced in the job

Appendix C: Data Dictionary for the Financial Sales Summary

This appendix contains selected details about each column in the financial sales summary spreadsheet. The spreadsheet is maintained on a monthly basis for cumulative actual amounts. Forecast amounts are entered annually so the forecast amounts must be compared to the actual amounts in the last month of the year. For other months, actual costs reflect the cumulative year to date but budget costs reflect the annual budgeted costs.

Note that the actual amounts are not just a summation of recognized revenue from the ERP database. The accounting department conducts a monthly reconciliation process to eliminate double counting of internal sales.

Financial Sales Summary

Column	Data Type	Definition
Summary_Sales_Id	LONG	Unique identifier for Financial Summary Sales
Actual_Units	INTEGER	Cumulative actual sales units until the ending date
Actual_Amount	CURRENCY	Cumulative actual sales amount until the ending date
Forecast_Units	INTEGER	Forecasted sales units for the year
Forecast_Amount	CURRENCY	Forecasted sales amount for the year
Location_Id	LONG	Identifier of the location
Sales_Class_Id	LONG	Identifier of the sales class
Begin_Date	DATE	Beginning date of actual units and amounts, usually the first day of the month
End_Date	DATE	Ending date of actual units and amounts, usually the last day of the month

Appendix D: Data Dictionary for the Financial Cost Summary

This appendix contains selected details about each column in the financial cost summary spreadsheet. The spreadsheet is maintained on a monthly basis for cumulative actual amounts. Budget amounts are entered annually so the forecast amounts must be compared to the actual amounts in the last month of the year for clear results. For other months, actual costs reflect the cumulative year to date but budget costs reflect the annual budgeted costs.

Note that the actual costs are not just a summation of costs from the ERP database. The accounting department conducts a monthly reconciliation process to eliminate double counting of actual costs as some shared costs for materials, overhead, and labor can be double counted.

Financial Cost Summary

Column	Data Type	Definition
Summary_Cost_Id	LONG	Unique identifier for Financial Summary Cost
Actual_Units	INTEGER	Cumulative actual units until the end date
Actual_Labor_Cost	CURRENCY	Cumulative actual labor costs until the end date
Actual_Material_Cost	CURRENCY	Cumulative actual material cost until the end date
Actual_Machine_Cost	CURRENCY	Cumulative actual machine cost until the end date
Actual_Overhead_Cost	CURRENCY	Cumulative actual overhead cost until the end date
Budget_Units	INTEGER	Annual budgeted units
Budget_Labor_Cost	CURRENCY	Annual budgeted labor cost
Budget_Material_Cost	CURRENCY	Annual budgeted material cost
Budget_Machine_Cost	CURRENCY	Annual budgeted machine cost
Budget_Overhead_Cost	CURRENCY	Annual budgeted overhead cost
Location_Id	LONG	Identifier of the location
Machine_Type_Id	LONG	Identifier of the machine type
Sales_Class_Id	LONG	Identifier of the sales class
Begin_Date	DATE	Begin date of actual costs, usually the first day of the month
End_Date	DATE	End date of actual costs, usually the last day of the month

Appendix E: Data Dictionary for the Lead File

The lead file is extracted from the Customer Relationship Management (CRM) system periodically. The identifiers for customers, locations, sales agents, and sales classes are sometimes inconsistent with the ERP tables as the ERP database do not have a convenient interface.

Lead File

Column	Data Type	Definition
Lead_Id	LONG	unique identifier for lead
Quote_Qty	INTEGER	number of items quoted for a lead
Quote_Price	CURRENCY	price per item quoted for a lead
Quote_Value	CURRENCY	dollar total quoted for a lead
Success	BOOLEAN	marks whether the lead turns into a job
PO_Number	LONG	purchase order number if the lead turns into a job
Created_Date	DATE	date the lead was generated
Cust_Id	LONG	identifier of the customer associated with the lead
Location_Id	LONG	identifier of the location associated with the lead
Sales_Agent_Id	LONG	identifier of the sales agent associated with the lead
Sales_Class_Id	LONG	identifier of the sales class associated with the lead