

Schema Integration

Techniques for Building the ODS

New York Oracle Users Group

General Spring Meeting

March 12, 2013

101 Murray Street - New York, NY 10007

General Discussion Topics

- Present Biography
- Schema Integration
- What is an ODS
- ODS Architecture
- ODS Roles
- Schema Integration Described
- The Process
- A simple Example
- Importance of Master Data
- Importance Of Data Quality Processes

Presenter Biography



Angelo R Bobak is data architecture professional and published author with over 25 years experience in Business Intelligence, Data Architecture, Data Modeling, Master Data Management, and Data Quality. Currently he is working at ATOS Origin NA as a Director/Senior Data Architect in the areas of Global Master Data Management, Data Integration and Data Quality. Past experience includes positions as an IT consultant, manager and data architect with companies such as Praxair, Avaya, Pepsi and several financial institutions on Wall Street such as Merrill Lynch, Bankers Trust and International Securities Exchange (ISE).

He is the author of several books in the areas of data modeling and distributed database design and is authoring a course on data integration for eLearningCurve:



Presentation Goals

- Leave with an understanding of the ODS architecture and its application.
- Leave with a high level understanding of what schema integration is.
- Leave with a high level understanding of the steps involved
- Leave with and understanding why Master Data is extremely important.
- Lastly, understand the importance of Data Quality.

Schema Integration

In today's modern business environment, corporate entities are constantly merging or splitting, internal divisions are sold to different companies, and new business lines are created in order to meet the challenges of difficult economic times. Business data integration is a complex problem that must be solved when organizations change or enhance their internal structures. New IT departments must be merged with old ones, and transactional, operational, and master data must be integrated in order to be managed efficiently, if the business is expected to grow and be profitable.

The goal of this presentation is to present a simple yet thorough process that describes the challenges of business data integration and the solutions to these challenges. It will show you how the application of a technique called "schema integration" addresses these challenges. Schema integration is both a theory and process that was pioneered by experts in the field of data management. We will discuss the techniques of two of these pioneers, M. Tamer Ozsü and Patrick Valduriez in the design of an Operational Data Store (ODS) for a small business.

M. Tamer Ozsü and Patrick Valduriez also discussed distributed database architectures and related topics such as distributed transaction processing and federated database architectures in their books and papers.

What is An ODS?

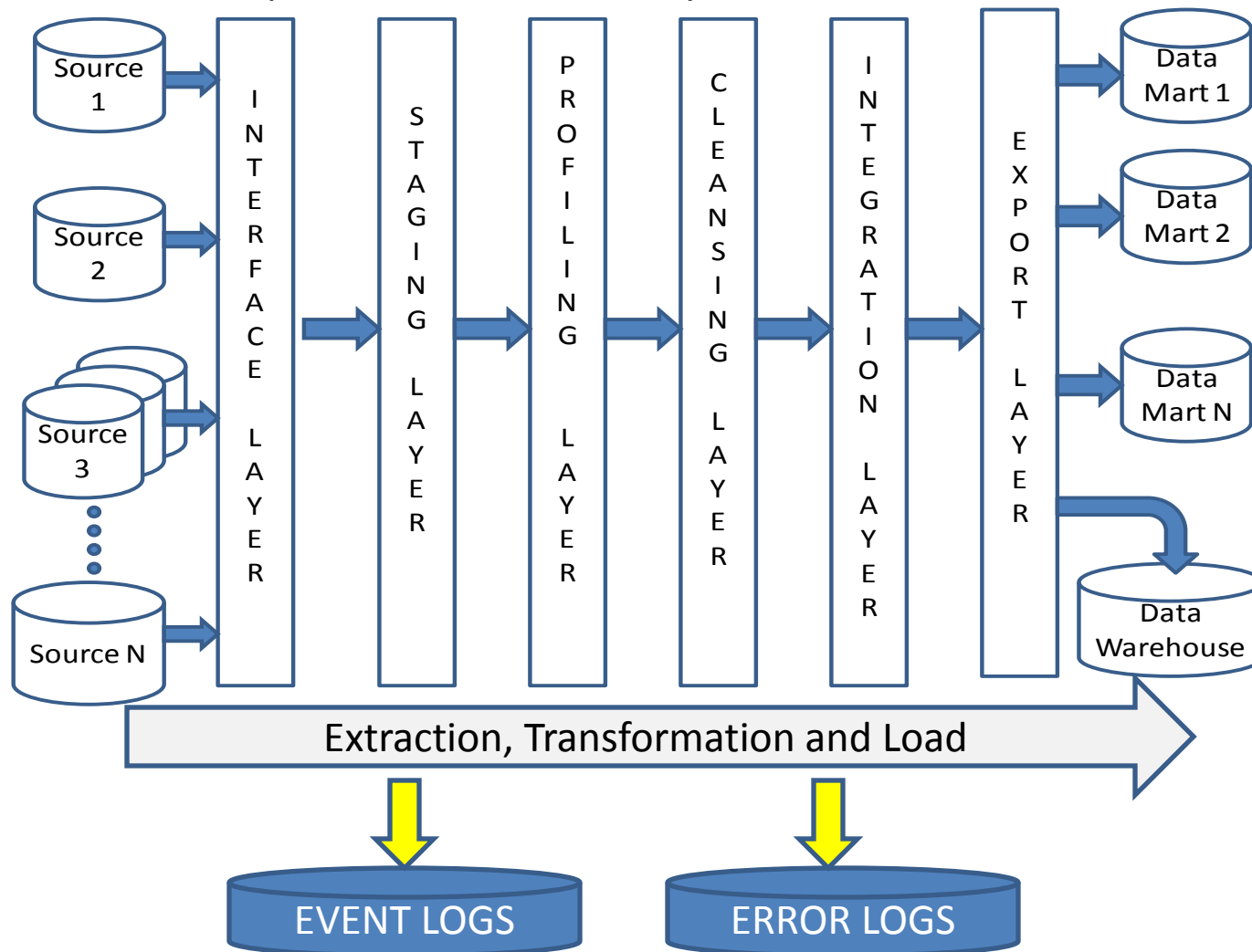
What is an Operational Data Store? This discussion provides some basic foundational concepts of the ODS: what it is, how it is used, and its role in a data warehouse architecture and data integration project. It also identifies some of the challenges faced when designing this data integration model. Specifically, we will address the various layers of the architecture:

- The Interface Layer
- The Data Staging Layer
- The Data Profiling Layer
- The Data Cleansing Layer
- The Data Integration Layer
- The Export Layer.

Each layer is separated by one or more ETL processes that loads, transfers, measures, cleanses and delivers the data of interest.

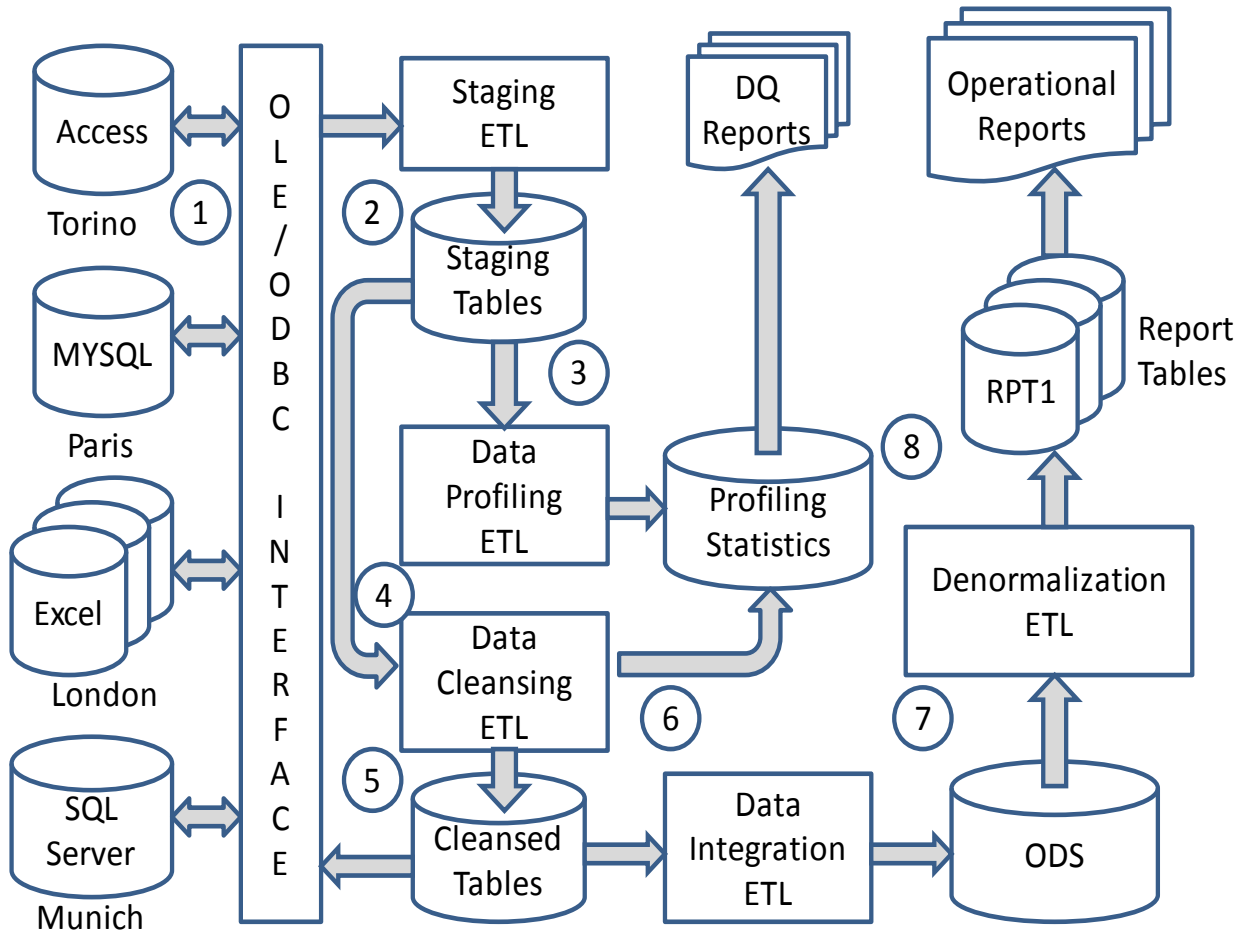
What is an Operational Data Store

Operational Data Store Layered Architecture



Each layer is separated by one or more ETL processes that loads, transfers, measures, cleanses and delivers the data of interest

ODS Detailed Architecture



DATE: 10/10/2012
DATA PROFILE REPORT

TRJ NAME	COL NAME	NULLS	DUPLICATES	OUT OF RANGE	MAX VALUE	MIN VALUE	DISTINCT VALUES
Date		0	0	0	1	00	01
	State Fy	1	0	0	AL	00	02
	State Code	1	0	0	ALABAMA	ALABAMA	02
	State Name	1	0	0	AL	AL	02
	Country Code	0	0	0	US	US	01

ROWS SAMPLED 11

Statstics Score	NULLS	DUPLICATES	OUT OF RANGE	MAX VALUE	MIN VALUE	DISTINCT VALUES
State Fy	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%
State Code	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%
State Name	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%
Country Code	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%

DATE: 10/10/2012
DATA PROFILE REPORT

TRJ NAME	COL NAME	NULLS	DUPLICATES	OUT OF RANGE	MAX VALUE	MIN VALUE	DISTINCT VALUES
Date		0	0	0	1	00	02
	State Fy	1	0	0	AL	00	02
	State Code	1	0	0	ALABAMA	ALABAMA	02
	State Name	1	0	0	AL	AL	02
	Country Code	0	0	0	US	US	01

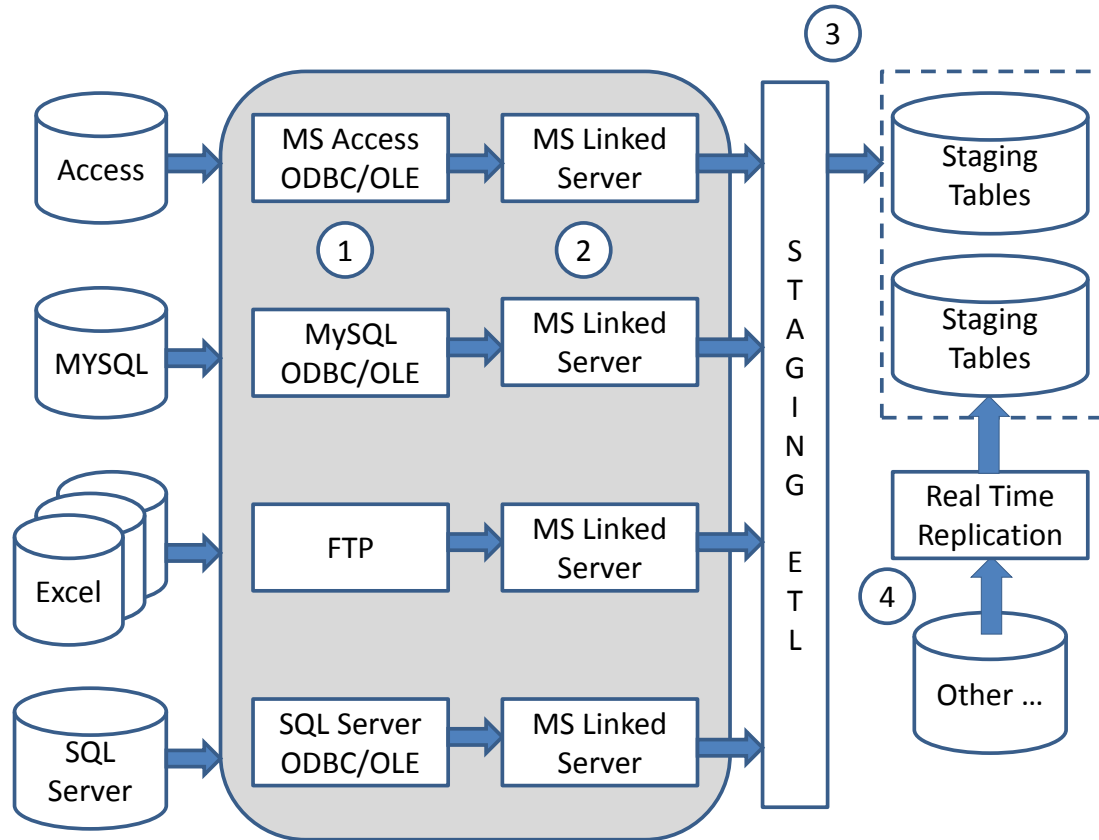
ROWS SAMPLED 00

Statstics Score	NULLS	DUPLICATES	OUT OF RANGE	MAX VALUE	MIN VALUE	DISTINCT VALUES
State Fy	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%
State Code	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%
State Name	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%
Country Code	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%

- Stage the data
- Profile the data
- Cleanse & enrich
- Feedback cleansed data
- Collect DQ statistics
- Merge the data
- Send to Consumers

Identify NULLS, duplicates, MAX/MIN values, unique values, out of bounds values, etc... and store in statistics tables for reporting and DQ scorecards

Typical Interfaces

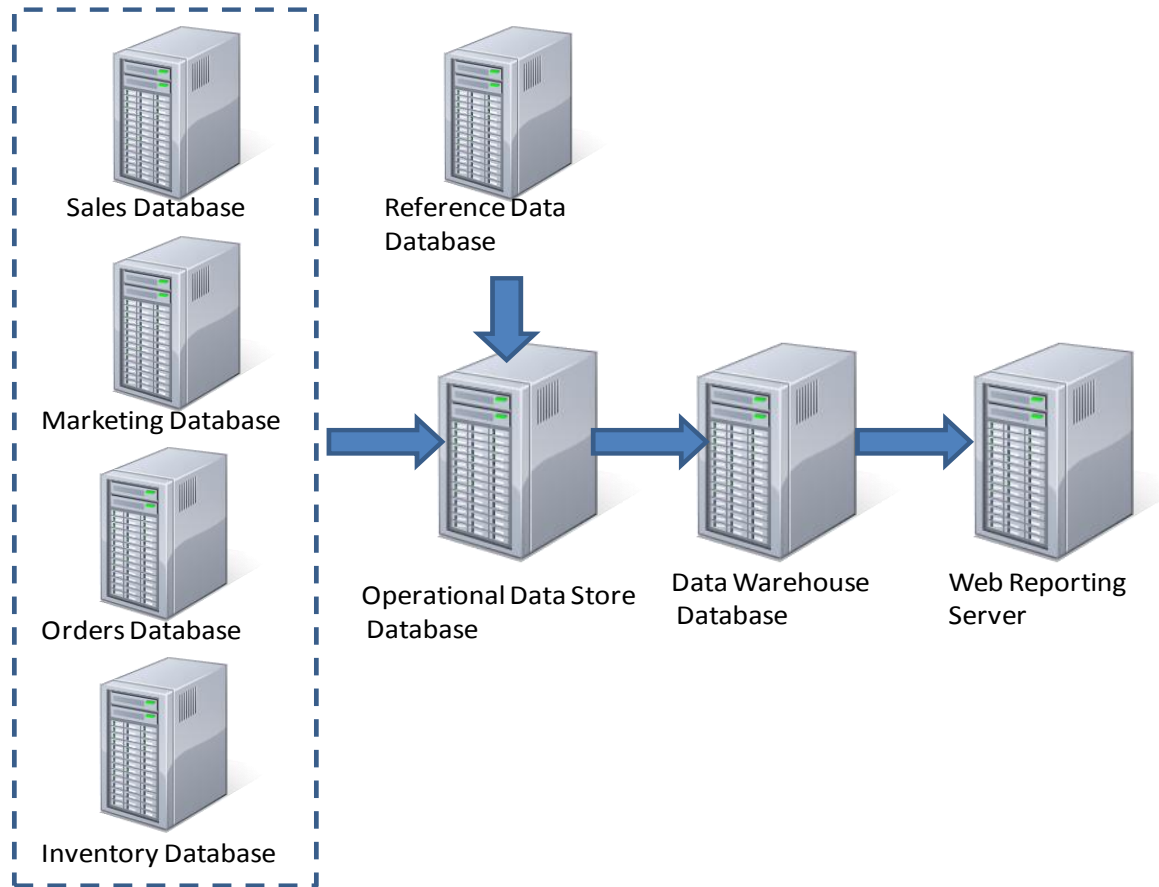


Standard industry interfaces based on popular data connectivity protocols for exchanging data between heterogeneous databases and data sources include:

- OLEDB - Object Linking and Embedding, Database
- ODBC Object Data Base Connectivity
- FTP File Transfer Protocol

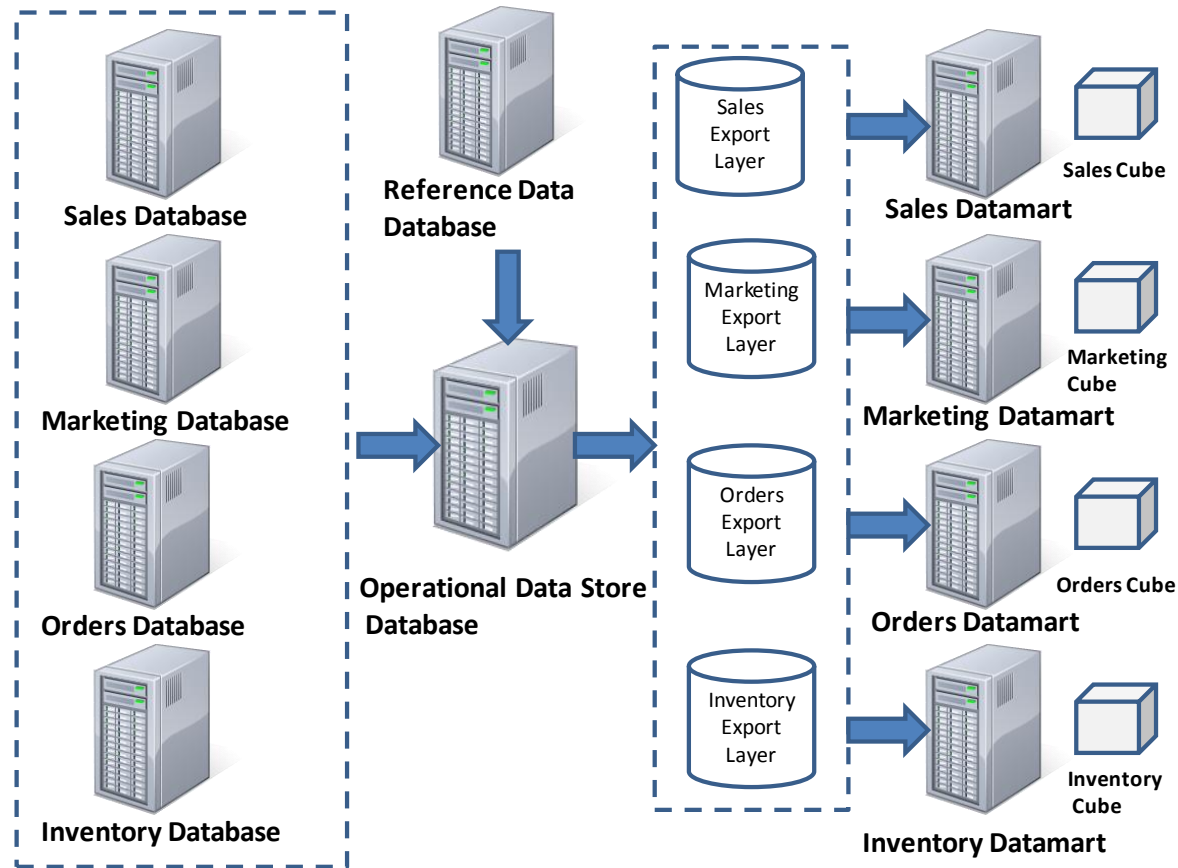
These and other protocols are used, in addition to real time replication data delivery interfaces. Of particular interest are the real time data replication interfaces that allow delivery of real time data from a source transactional database to a target database. ODS architects are interested in transactional replication because it allows any data warehouse to be kept up to date with accurate data for reporting and OLAP analysis in near or real time.

ODS Roles – supporting Data Warehouse



Data warehouse architectures are repositories of historical data, specially formatted to supply snapshots in time for analysis of the data. Analysis is performed by looking at the data in a multi-dimensional manner. That is, from multiple perspectives. Additionally, users can navigate up and down hierarchies so as to aggregate and decompose aggregated statistics all the way down to the transactional level if needed.

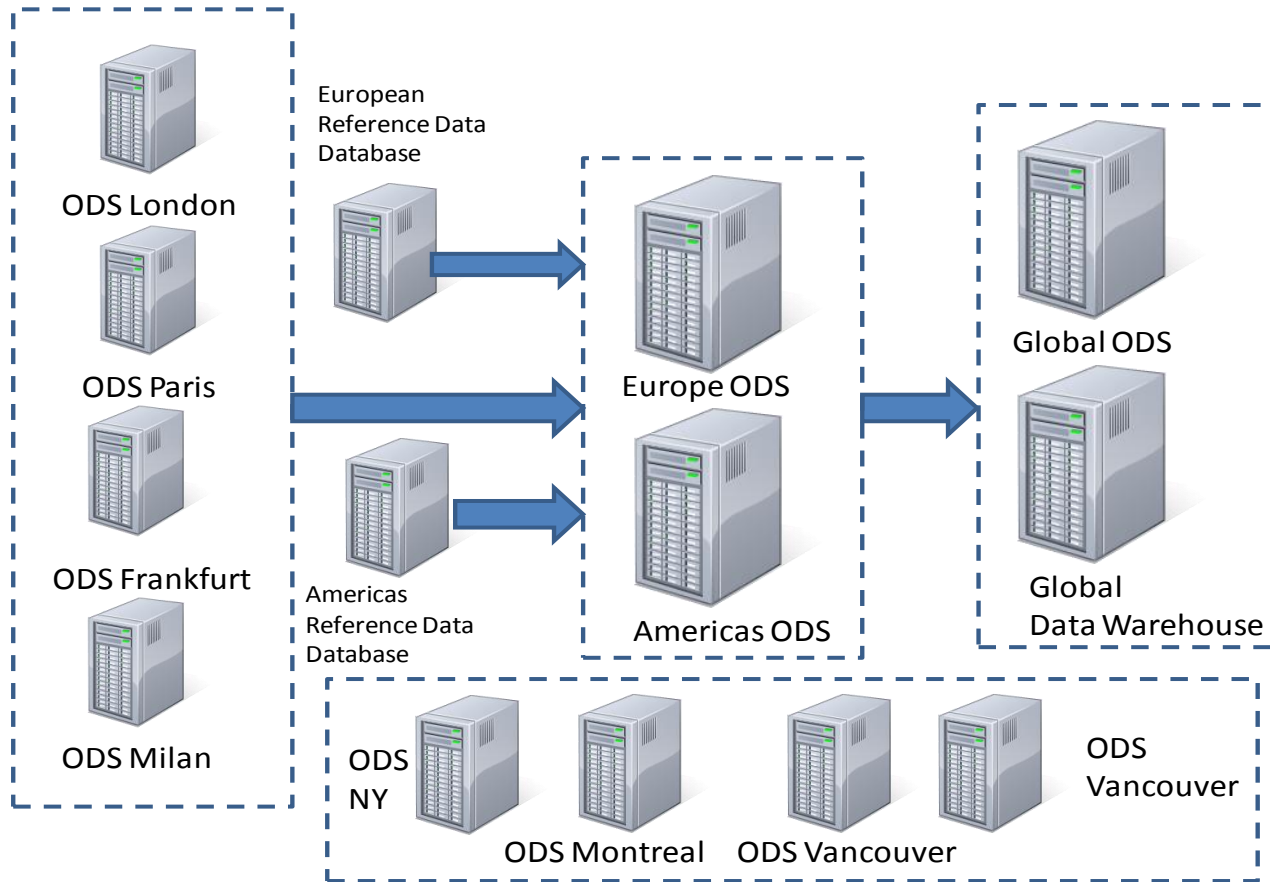
ODS Roles – supporting Data Marts



The basic functionality in this scenario is the same as that of the data warehouse architecture we discussed in the last section. Operational data is pulled, staged, profiled and cleansed. Reference data is used to clean up missing values or resolve issues like duplicate names.

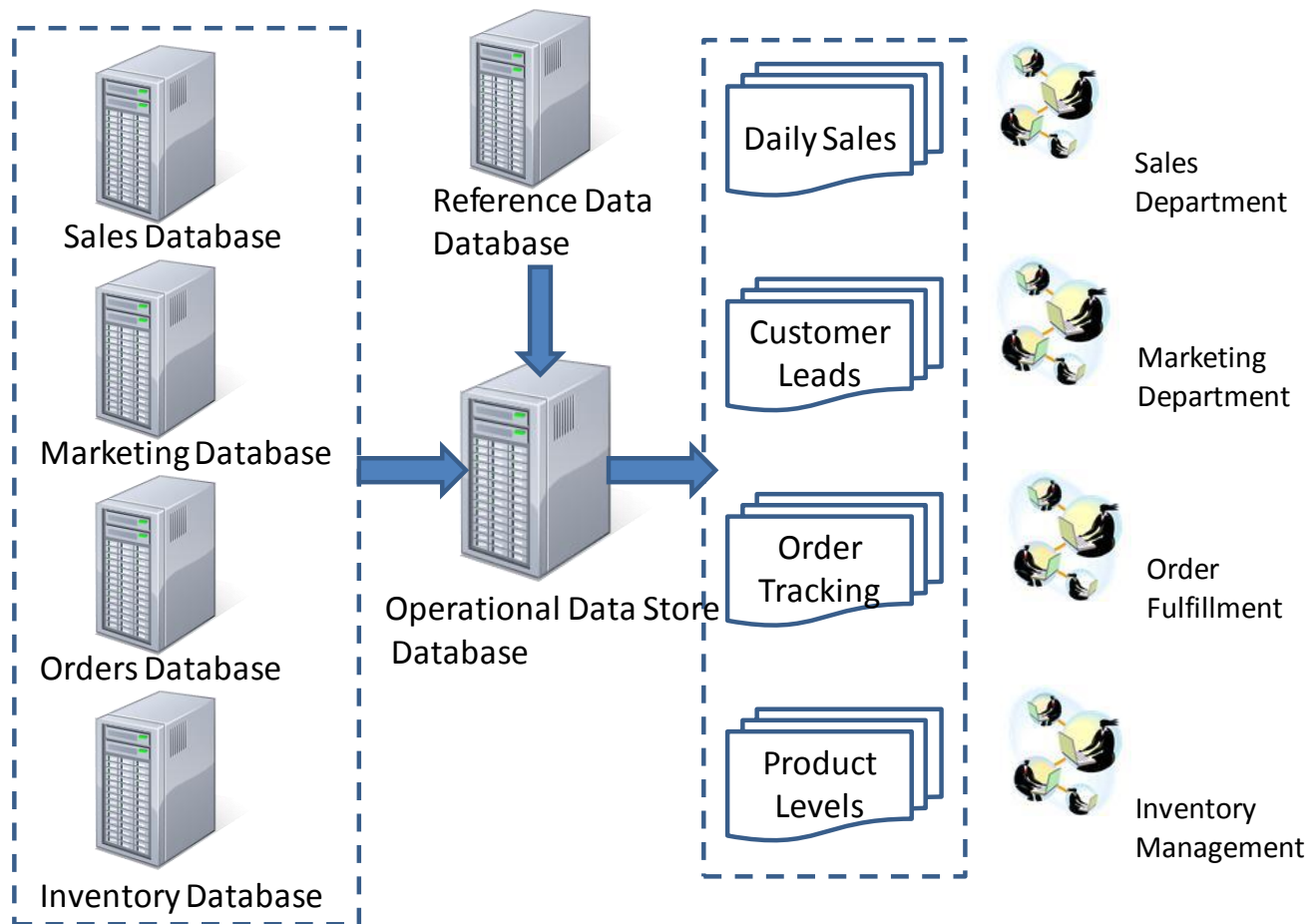
The ODS then exports the data by subject area. By export we mean that it stages the data in a dedicated layer of tables and views. For example, after profiling and cleansing the sales data, it exposes it to the sales data mart. This technique adds a security layer so that only dedicated sales analysts and financial types can examine the data.

Member of Distributed Architecture



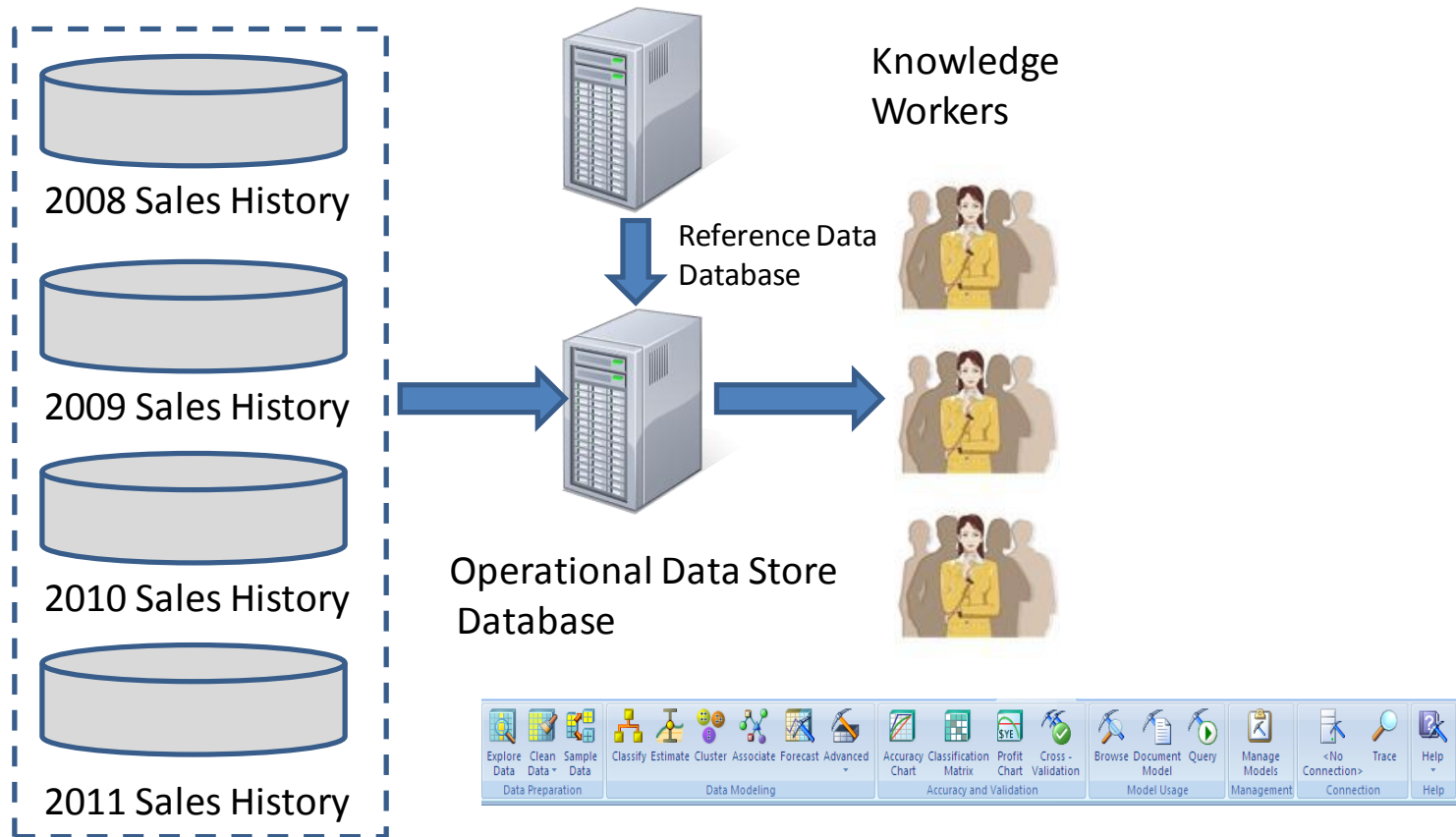
In this scenario, a series of ODS exist in Europe and in the Americas, namely the United States and Canada. Each ODS collects data from operational systems in its respective city. A regional reference database exists so as to capture customer, product and location information.

Source for Operational Reporting



The next role that we discuss supports the bread and butter reporting for any organization, namely operational reporting. Operational reporting is performed on a daily basis by various levels of business users in order to monitor the daily performance of the organization.

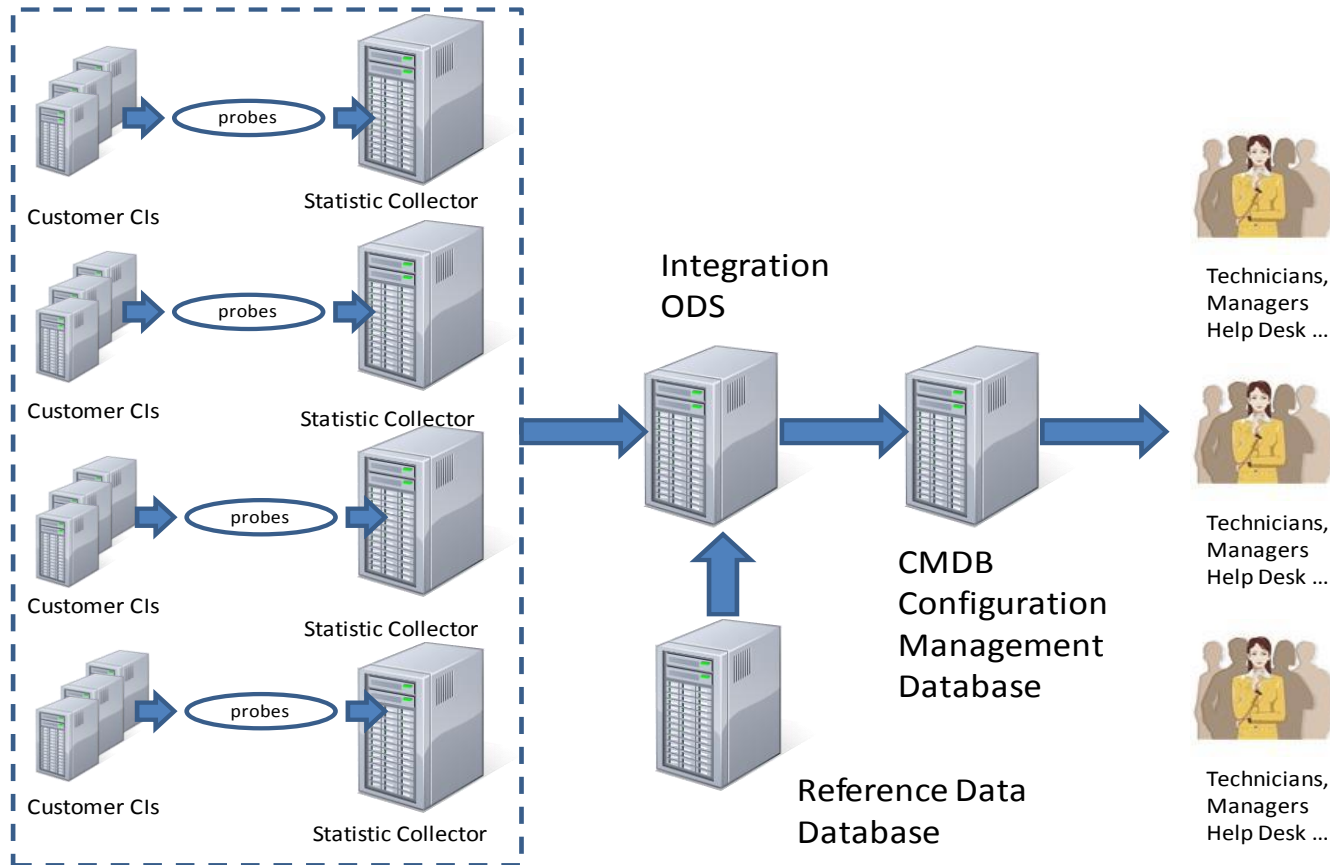
ODS – Providing Data Mining Support



Data mining is the process whereby interesting patterns of data are identified and visualized so that we can predict the future based on events of the past. Additionally, we can look at patterns so as to identify the causes of a result. For example, why do beer purchases usually appear with diaper purchases at super markets? (I didn't make this one up!)

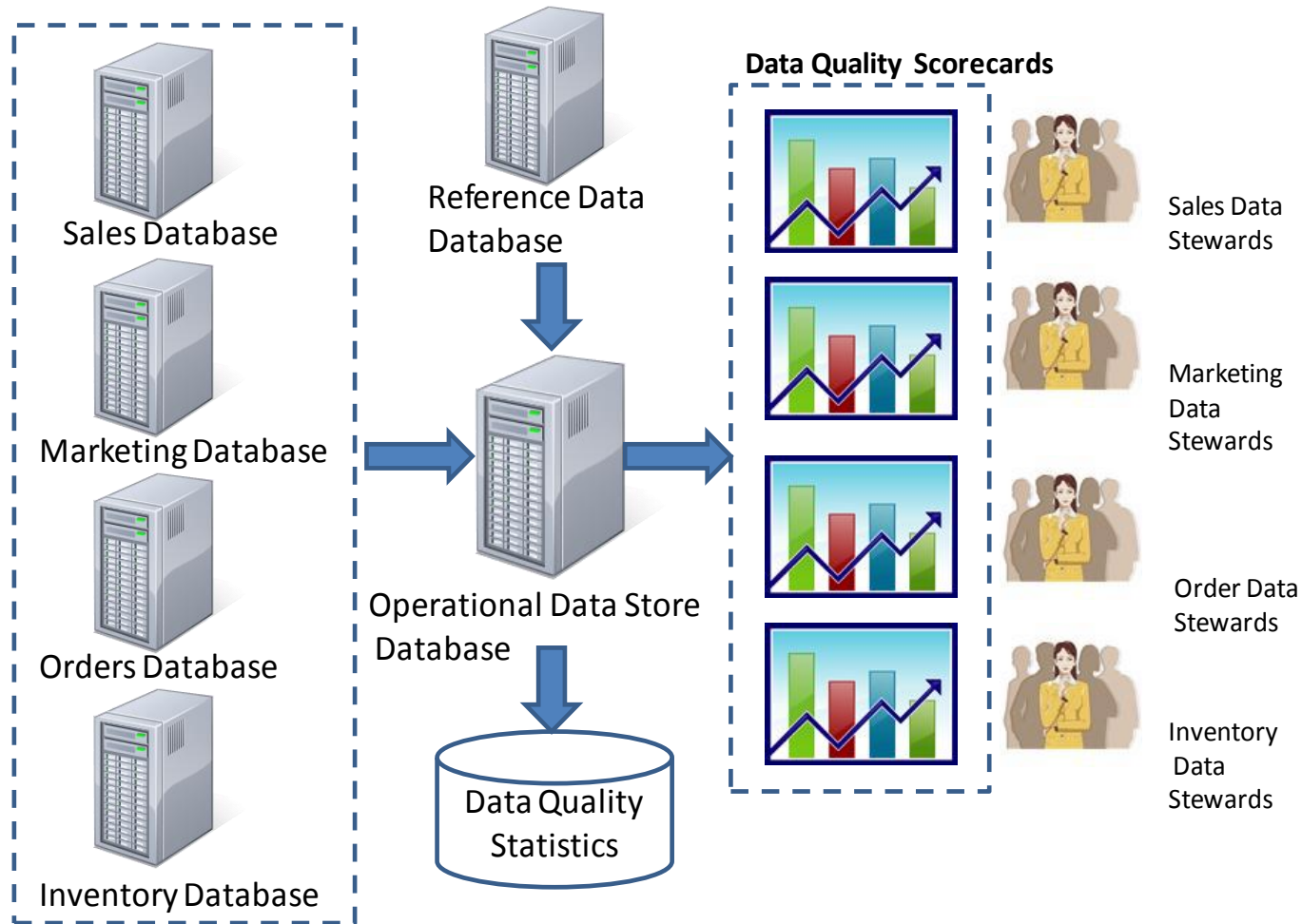
Below we see the ODS as an integrator of sales data so that it can be presented to a knowledge worker to predict sales patterns.

ODS – ITIL Support



One can consider the CMDB as a specialized data warehouse that stores configuration items and related reference data. A configuration item could be a server, memory in the server, CPU in the server, network information, routers, switches, racks, not to mention computer rooms, software and other items that are part of an IT infrastructure.

ODS – Data Quality Support



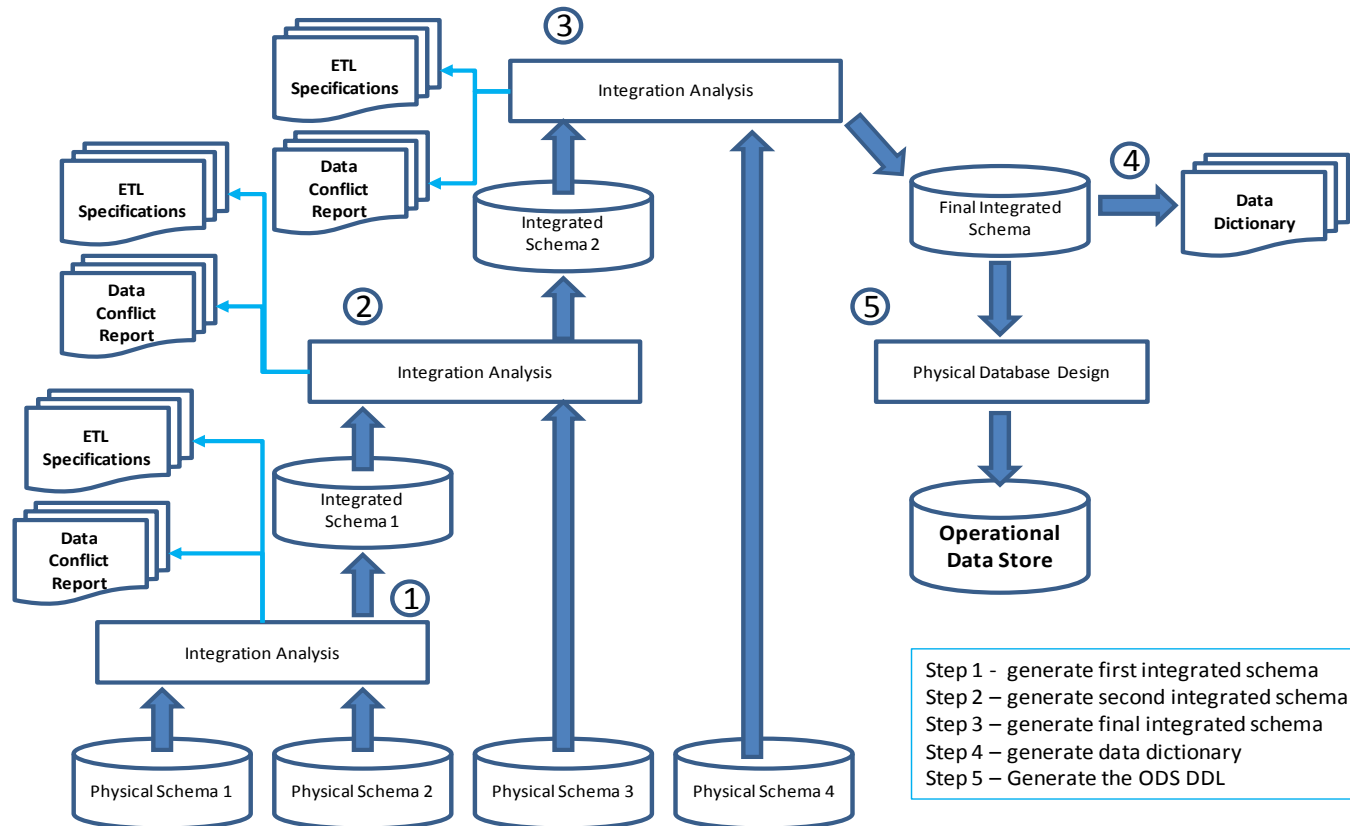
Lastly we look at the ODS as a data quality platform. The ODS coupled with a reference database provides any reporting or data integration architecture with a valuable set of processes to measure and cleanse data. Figure 4.7 shows the ODS, together with a reference database and a set of data quality tables.

What is Schema Integration?

A high level description of schema integration is the identification of common and uncommon or unique data objects in each schema. The uncommon data objects can stand on their own so we need not worry about them. It is the common objects we care about, specifically table keys, column and relationships that are common in each schema pair. This is what we will be integrating!

What is Schema Integration?

Binary Schema Integration – Technique 1

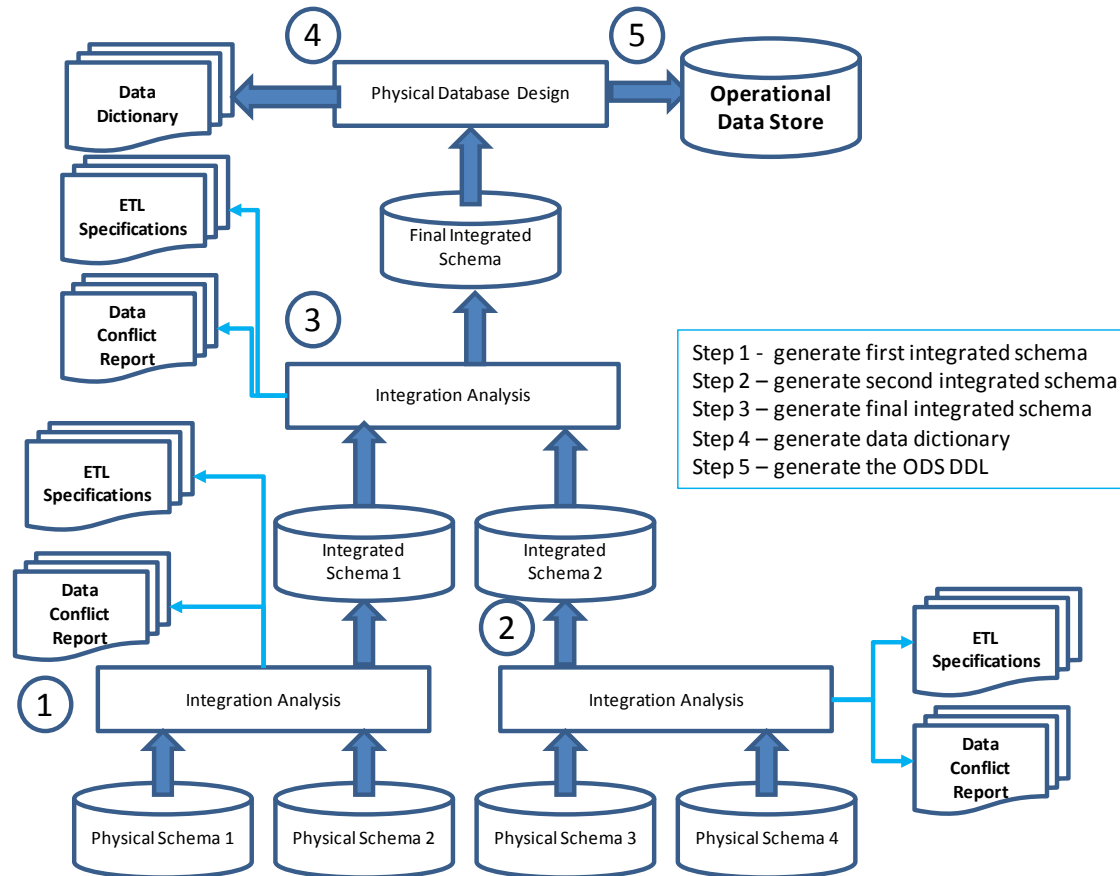


- Step 1 - generate first integrated schema
- Step 2 - generate second integrated schema
- Step 3 - generate final integrated schema
- Step 4 - generate data dictionary
- Step 5 - Generate the ODS DDL

- Step 1 – generate the first integrated schema
- Step 2 – generate the second integrated schema
- Step 3 – generate the final integrated schema
- Step 4 – generate the integrated schema data dictionary
- Step 5 – Generate the ODS DDL (Data Declaration Language).

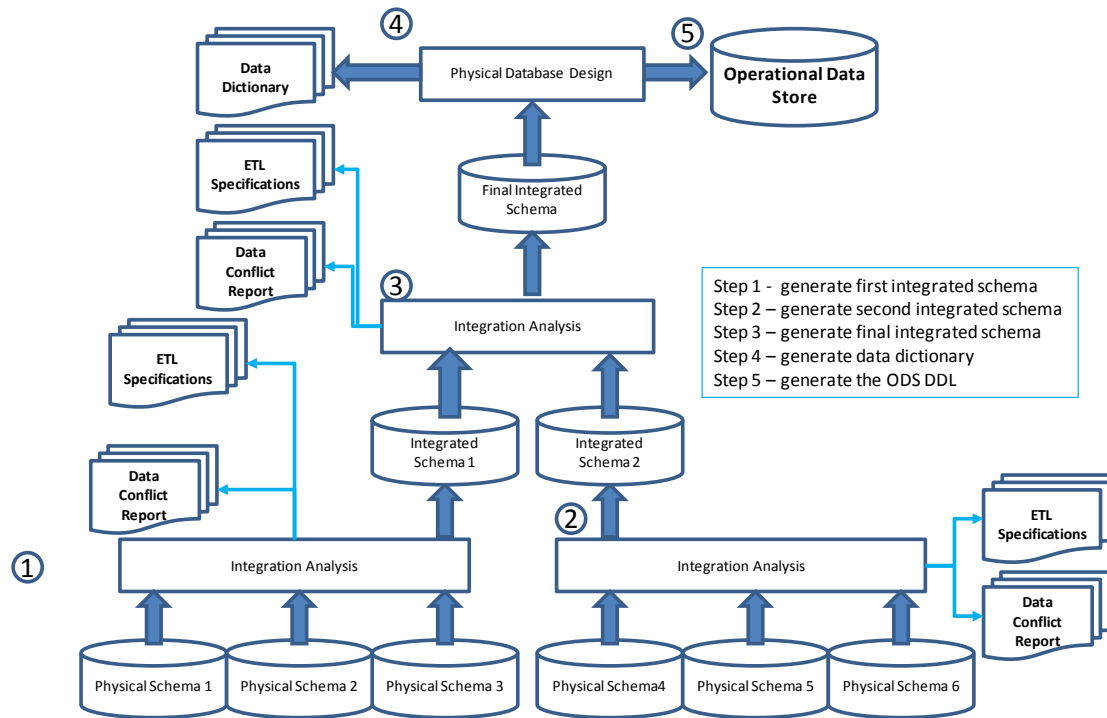
Binary Schema Integration, by its very name implies taking two source schemas at a time and combining them to generate one integrated schema.

Binary Schema Integration – Technique 2



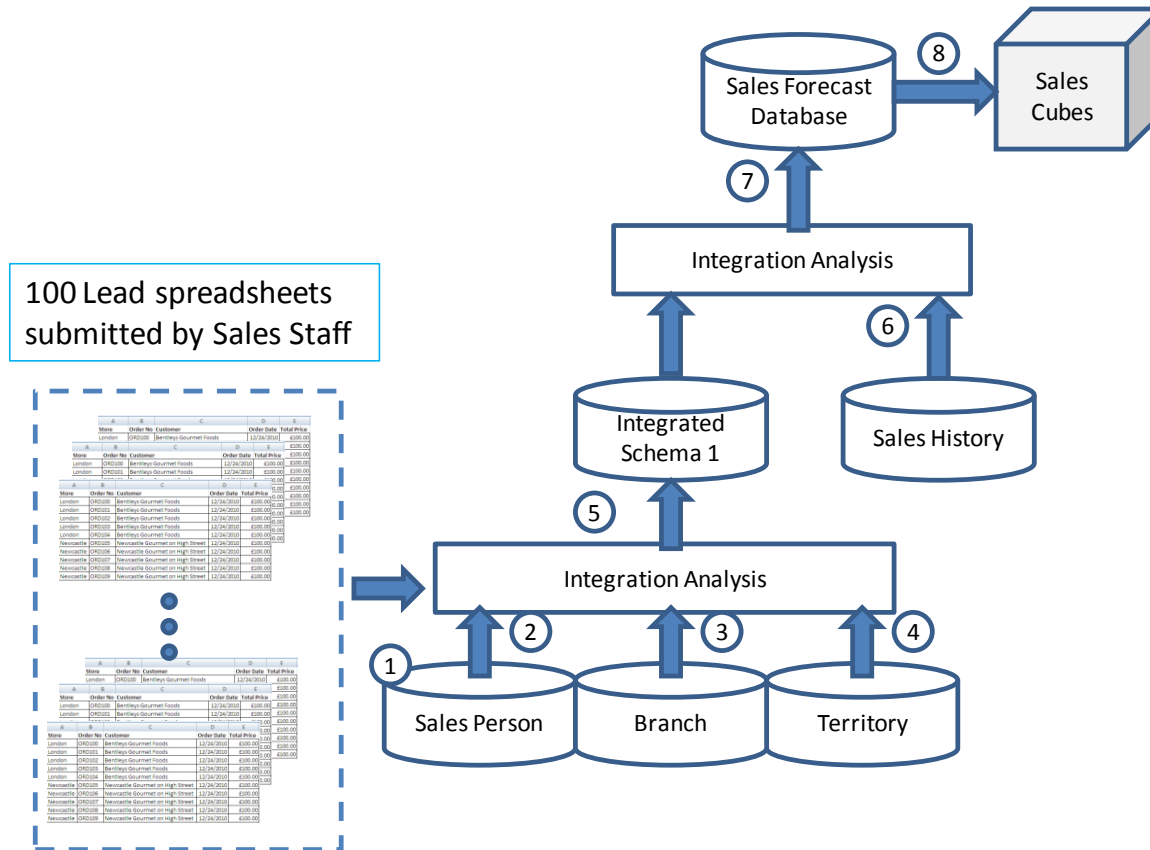
The first schema integration technique we discussed in the prior section took two source schemas, integrated them so as to produce an intermediate schema and generated a set of support documents, data dictionaries and ETL specifications. The key word here is “two”. This is why the inventors of this process called it Binary Schema Integration (BSI). The process continued until all source schemas were exhausted and a final integrated schema was created.

Tertiary Schema Integration



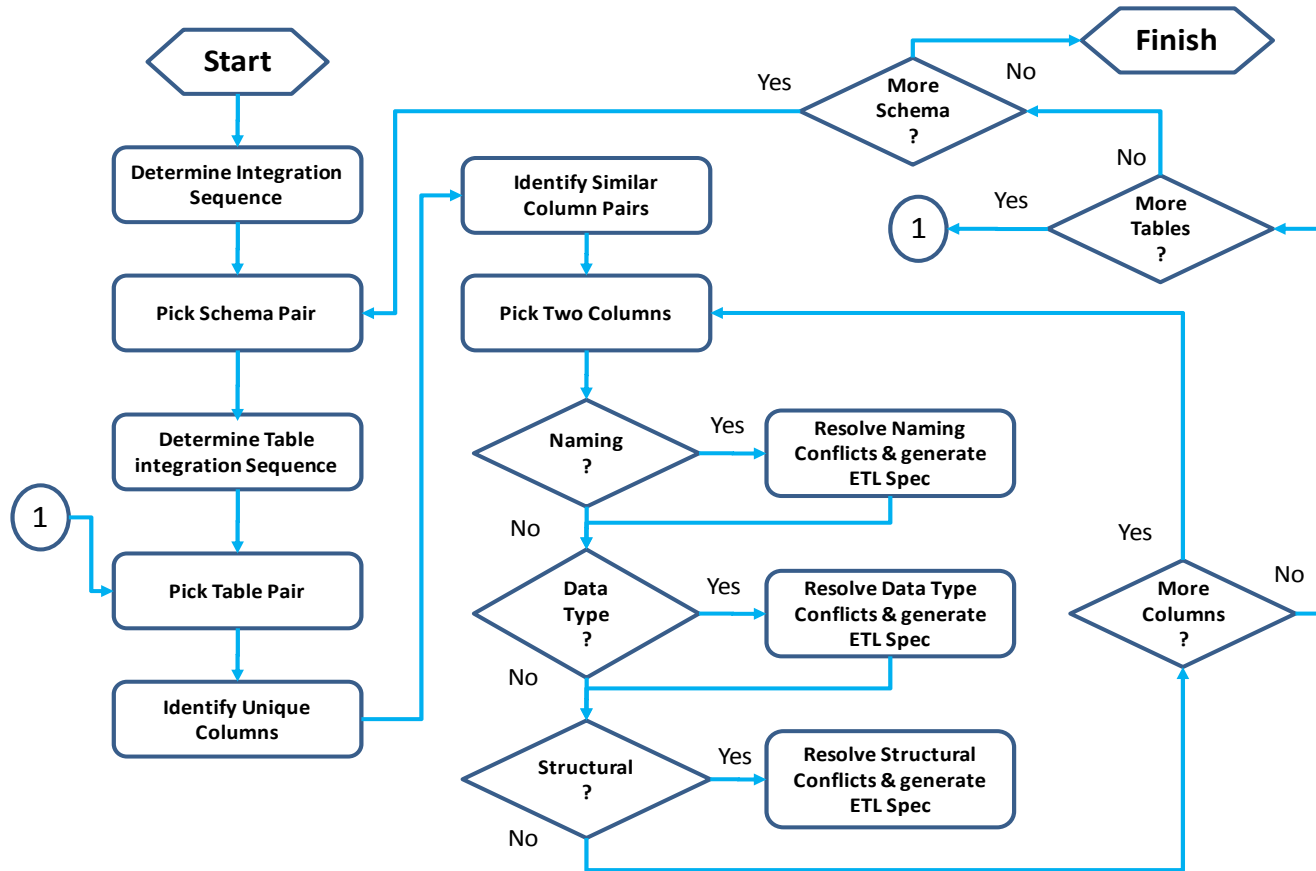
Another integration scheme is called tertiary schema integration which takes three schemas at a time as illustrated in the figure above.

N-ary Schema Integration



Now that we have examined the different types of schema integration techniques we are ready to examine the actual steps and what their inputs and outputs are.

Schema Integration – the Process

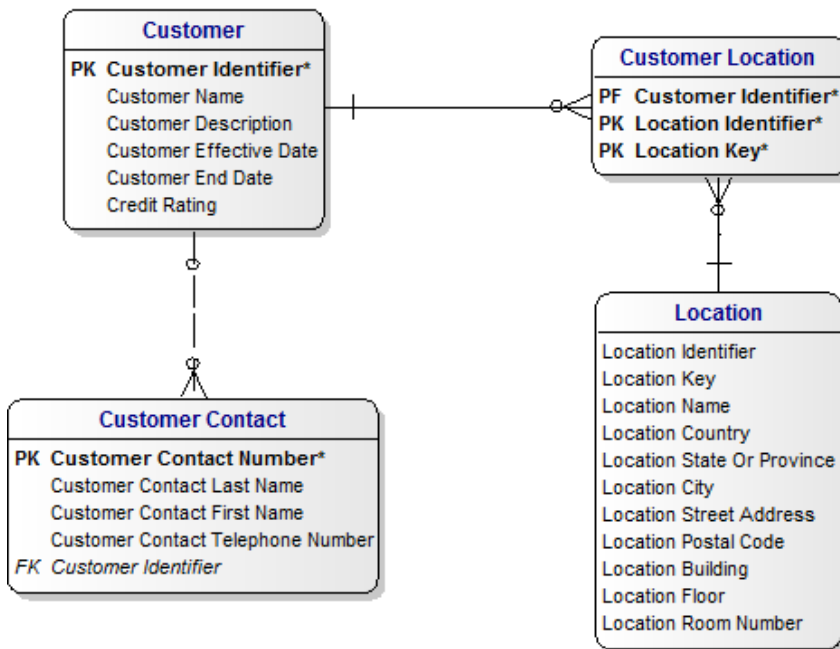


As you can see, it is an iterative process. That is, we repeat the steps until we have no more schemas to integrate.

A simple Customer ODS Model

Below we present a partial view of the final integrated data model for an ODS:

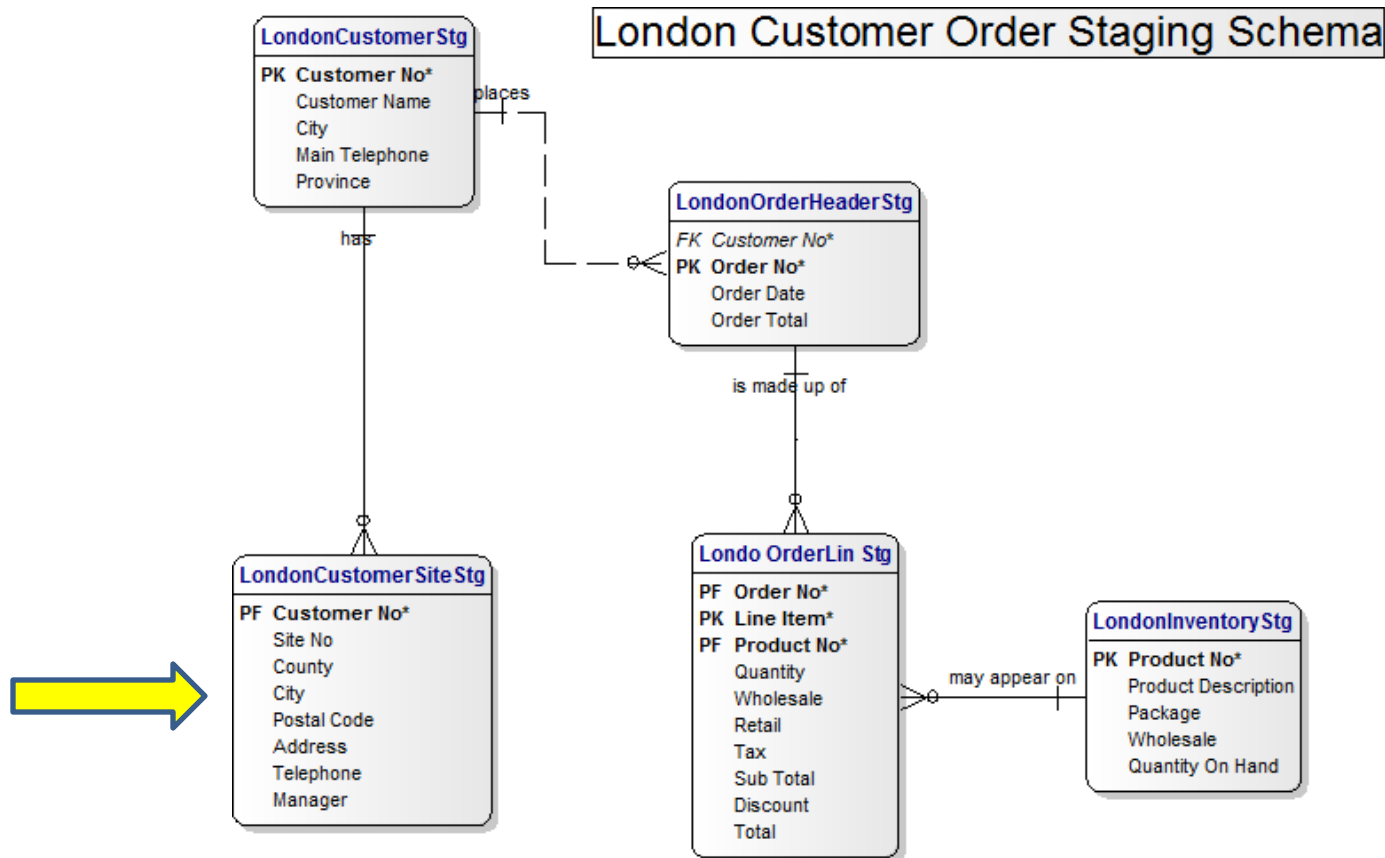
Customer and Customer Location
Final Integrated Model



- A customer can have zero, one or more customer locations.
- A location identifies the physical address of zero, one or more locations.
- A customer can have zero, one or more contacts.

This is a simple model but it supports the basic information one would require to deliver integrated customer data.

The first Customer Source Schema



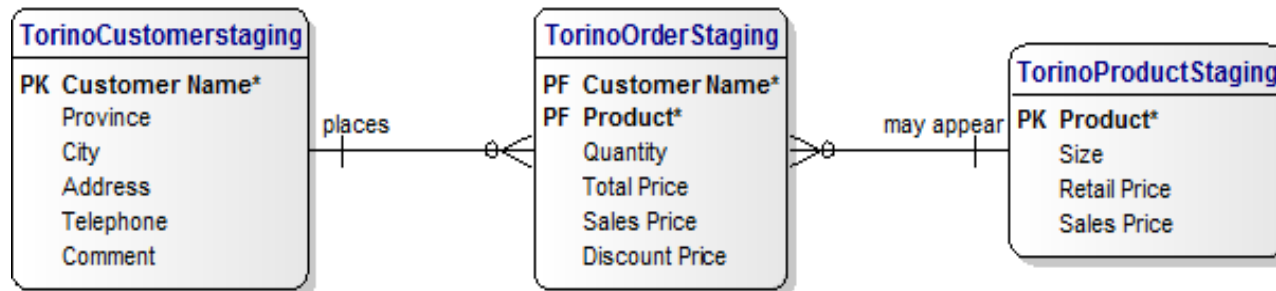
Here is the first schema to integrate, notice that there are only two customer related tables. Address information is located in the second table.

The Second Customer Source Schema



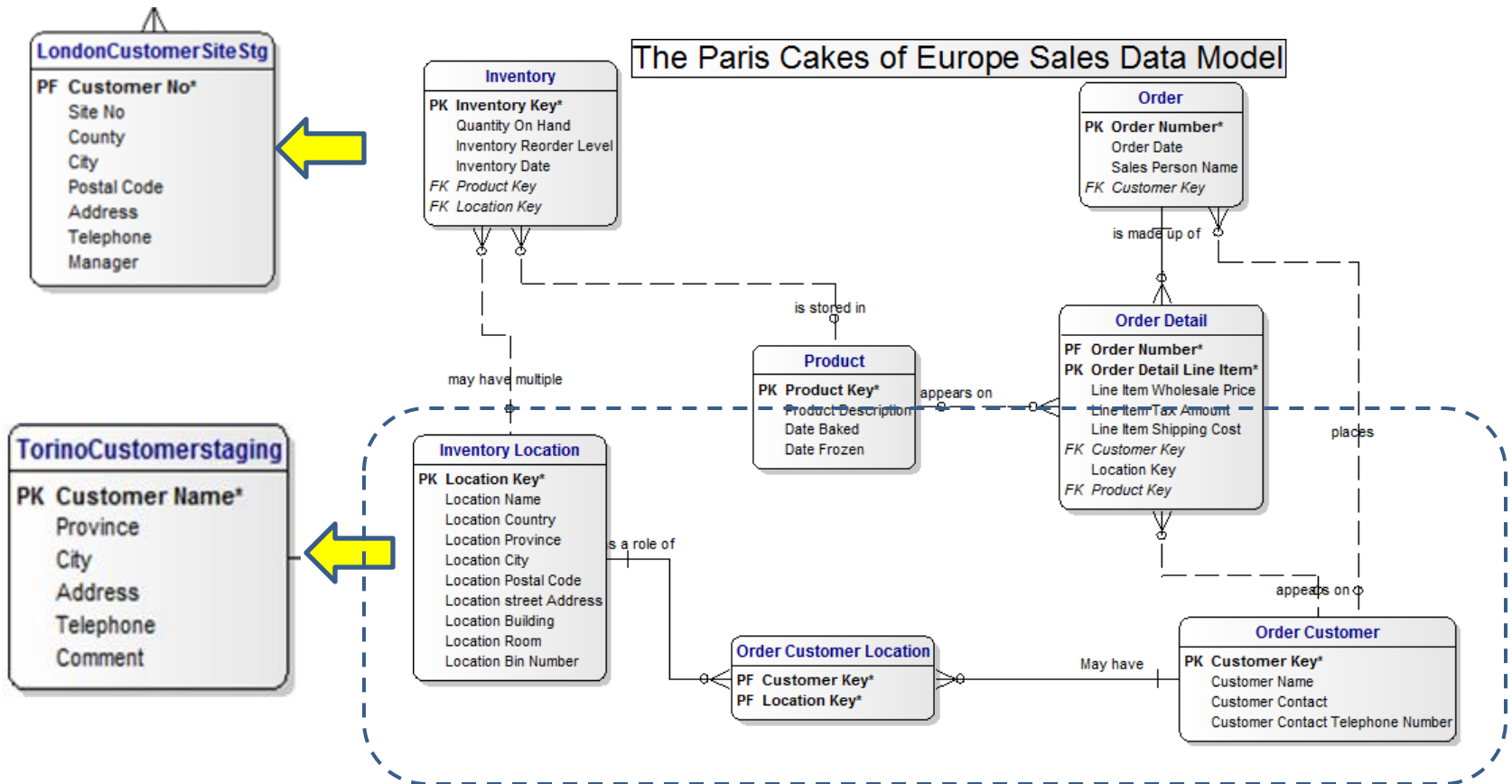
Prior customer table for reference.

Torino Order Staging Database



Here is the second schema to integrate, again only one customer related table. This time the address information is included in the customer table.

A Third Customer Source Schema

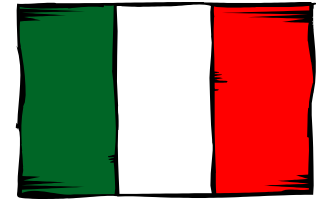
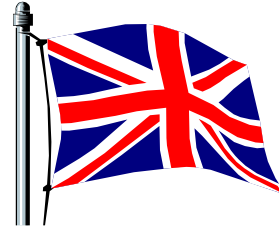
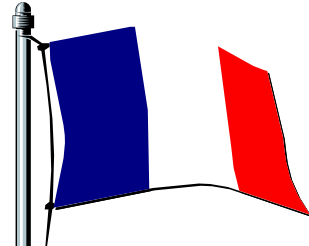


And here is the third schema to integrate, again only one customer related table. Notice the structural conflicts in that address columns are stored in different tables in two of the schema.

Integration Example

- Picking the Integration Sequence
- Data Conflicts
 - Data Type
 - Naming (antonyms and synonyms)
 - Structural
- Sample Data Conflicts Specification
- Sample ETL Specification

Picking the Integration Sequence



Source Schema	Location
Marketing DB	London
Customer Master	Paris
Product Master	New York
Order Master	Munich
Inventory Database	Paris
Product Master	New York
Returns	Munich

Marketing DB

Leads
SalesPerson
Product
Customer
Location

Customer Master

Customer
Customer Location
Customer Site
Customer Type
Address

Customer Master

Customer
Customer Location
Customer Site
Customer Type
Address

Order Master

Order Header
Order Line Item
Product
Product Type
Sales Person

Inventory Database

Inventory
Inventory Location
Address
Product

Product Master

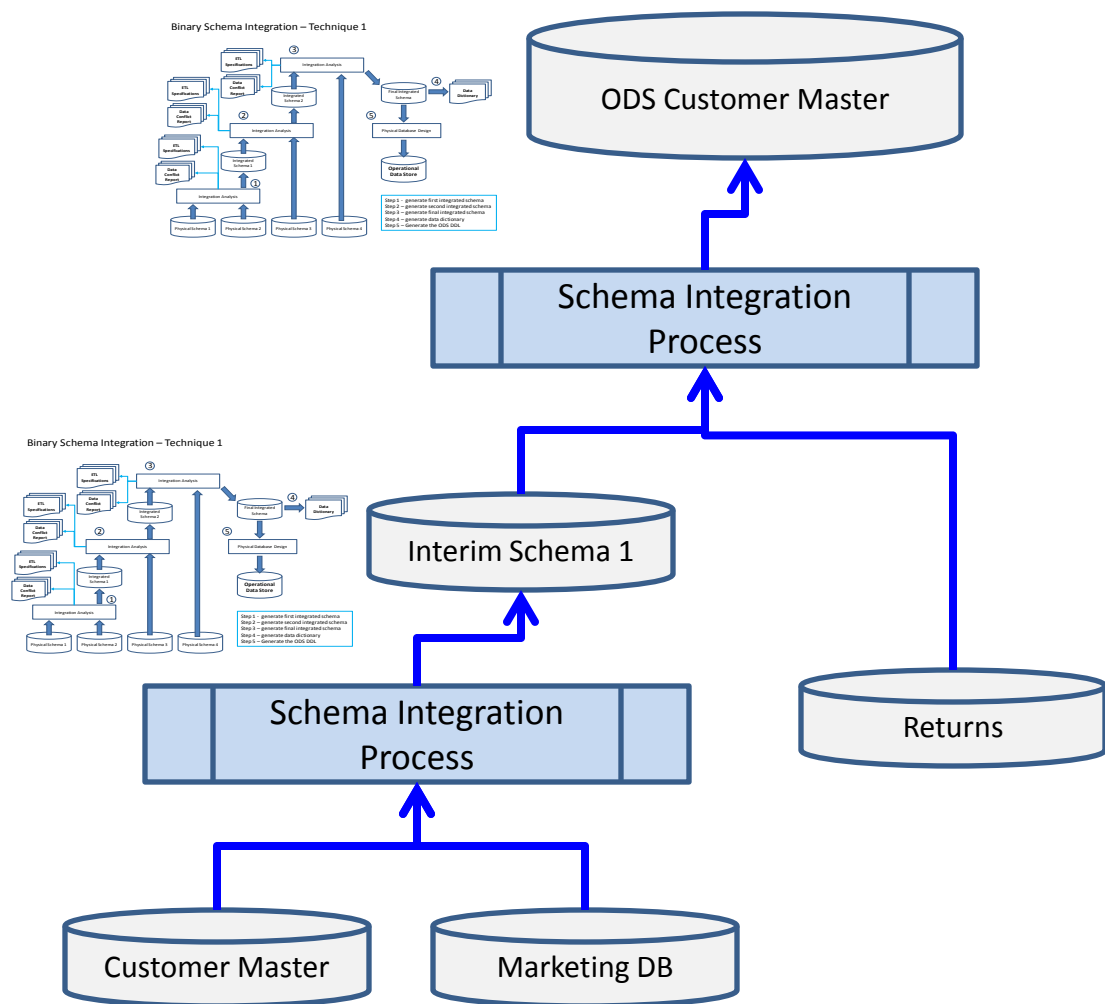
Product
Product Type
Product SubType
Product Price
Schedule
Product Details
Manufacturer
Vendor
Shipper

Returns

Return Header
Return Line Item
Customer
Sales Person

Pick Integration Sequence

Schema	Table
Customer Master	Address
Inventory Database	Address
Marketing DB	Customer
Customer Master	Customer
Returns	Customer
Customer Master	Customer Location
Customer Master	Customer Site
Customer Master	Customer Type
Inventory Database	Inventory
Inventory Database	Inventory Location
Marketing DB	Leads
Marketing DB	Location
Product Master	Manufacturer
Order Master	Order Header
Order Master	Order Line Item
Product Master	Price Schedule
Marketing DB	Product
Product Master	Product
Order Master	Product
Inventory Database	Product
Product Master	Product
Product Master	Product Details
Product Master	Product Price Schedule
Product Master	Product Sub Type
Product Master	Product SubType
Product Master	Product Type
Order Master	Product Type
Product Master	Product Type
Returns	Return Header
Returns	Return Line Item
Order Master	Sales Person
Returns	Sales Person
Marketing DB	SalesPerson
Product Master	Shipper
Product Master	Vendor



Project stake holders recognize that a big issue and immediate requirements is to create a customer master in order to manage new and existing customers.

Pick Table Integration Sequence


Schema	Table
Customer Master	Address
Inventory Database	Address
Marketing DB	Customer
Customer Master	Customer
Returns	Customer
Customer Master	Customer Location
Customer Master	Customer Site
Customer Master	Customer Type

Common Tables

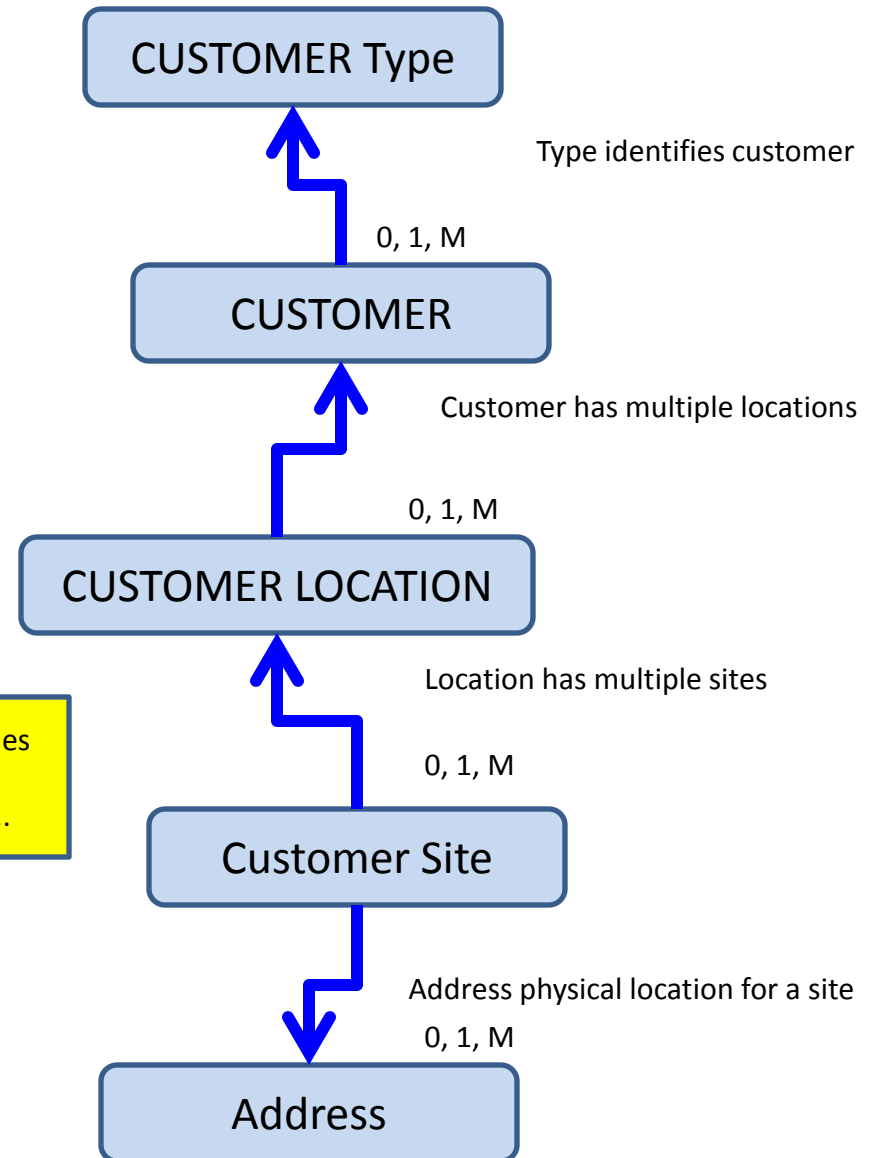
Schema	Table
Marketing DB	Customer
Customer Master	Customer
Returns	Customer

Unique Tables

Schema	Table
Customer Master	Address
Customer Master	Customer Location
Customer Master	Customer Site
Customer Master	Customer Type

Note:
 Child  Parent
 0, 1, M

These are the ones we want to concentrate on....



Pick Column Integration Sequence (1)

Schema	Table	Column	Data Type	Length	Business Name
Returns	Customer	CUST_ID	NUMBER	4	Customer Identifier
Returns	Customer	CUST_NAME	CHARACTER	128	Customer Name
Returns	Customer	EFFECTIVE_DATE	DATE	10	Customer Effective Date
Returns	Customer	END_DATE	DATE	10	Customer End Date
Returns	Customer	COUNTRY	CHARACTER	3	Country Code
Returns	Customer	PROVINCE	CHARACTER	64	Province
Returns	Customer	ADDRESS	CHARACTER	256	Address
Returns	Customer	POSTAL_CODE	CHARACTER	32	Postal Code

Schema	Table	Column	Data Type	Length	Business Name
Customer Master	Customer	CID	CHARACTER	6	Customer Identifier
Customer Master	Customer	CACCT_NO	CHARACTER	12	Customer Account No
Customer Master	Customer	CNAME	CHARACTER	64	Customer Name
Customer Master	Customer	CTYPE	CHARACTER	3	Customer Type

Adding logical business names helps us sort the columns to see similar related data. (Next slide)

Pick Column Integration Sequence(2)

Data Naming Synonym Conflict
This is actually the account number and not a sequential surrogate key identifier

	Table	Column	Data Type	Length	Business Name
	Customer	ADDRESS	CHARACTER	256	Address
	Customer	COUNTRY	CHARACTER	3	Country Code
Customer Master	Customer	CACCT_NO	CHARACTER		Account No
Returns	Customer	EFFECTIVE_DATE	DATE	10	Customer Effective Date
Returns	Customer	END_DATE	DATE	10	Customer End Date
Returns	Customer	CUST_ID	NUMBER	4	Customer Identifier
Customer Master	Customer	CID	CHARACTER	6	Customer Identifier
Returns	Customer	CUST_NAME	CHARACTER	128	Customer Name
Customer Master	Customer	CNAME	CHARACTER	64	Customer Name
Customer Master	Customer	CTYPE	CHARACTER	3	Customer
Returns	Customer	POSTAL_CODE	CHARACTER	32	Postal Code
Returns	Customer	PROVINCE	CHARACTER	64	Province

Data Type Conflict

CUST_0001
CUST_0002
CUST_0003
CUST_0004

Versus

1000
1001
1002
1003
1004

Data Len Conflict

Data Naming Conflict

Data and Structural conflicts really stand out now....

Example of a Structural Conflict

Schema	Table	Column	Data Type	Length	Business Name
Returns	Customer	CUST_ID	NUMBER	4	Customer Identifier
Returns	Customer	CUST_NAME	CHARACTER	128	Customer Name
Returns	Customer	EFFECTIVE_DATE	DATE	10	Customer Effective Date
Returns	Customer	END_DATE	DATE	10	Customer End Date
Returns	Customer	COUNTRY	CHARACTER	3	Country Code
Returns	Customer	PROVINCE	CHARACTER	64	Province
Returns	Customer	ADDRESS	CHARACTER	256	Address
Returns	Customer	POSTAL_CODE	CHARACTER	32	Postal Code

Conflict	Resolution
Structural	ETL_SPEC_0001
Structural	ETL_SPEC_0001
Structural	ETL_SPEC_0001
Structural	ETL_SPEC_0001

Schema	Table	Column	Data Type	Length	Business Name
Customer Master	Address	Address Key	NUMBER	4	Address Identifier Key
Customer Master	Address	Country Code	CHARACTER	2	ISO 2 Character Country Code
Customer Master	Address	Province	CHARACTER	64	Province Name
Customer Master	Address	State	CHARACTER	128	State Name
Customer Master	Address	Address 1	CHARACTER	256	Street Address 1
Customer Master	Address	Address 2	CHARACTER	256	Street Address 2
Customer Master	Address	Building	CHARACTER	32	Building Name
Customer Master	Address	Floor	CHARACTER	32	Floor Number
Customer Master	Address	Room Number	CHARACTER	8	Room Number
Customer Master	Address	Postal Code	CHARACTER	12	Postal Code

ETL SPECIFICATION

ETL_SPEC_0001	Break out and merge with Address entity
---------------	---



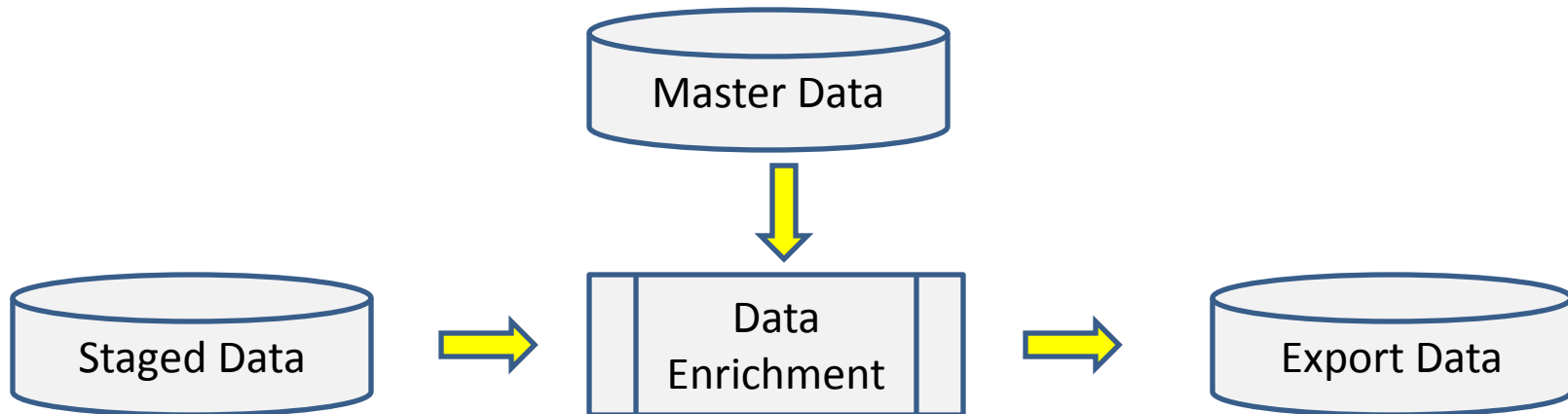
Design a template that maps the data to the conflicts, a specification number for the conflicts and then the specification itself. The specification can be enhanced with other documents, like source to target spreadsheets, data enrichment flow charts and process diagrams...

Importance of Master Data

COUNTRY	CALLING CODE	INTERNATIONAL DIALING PREFIX
Afghanistan	93	0
Albania	355	0
Algeria	213	0
American Samoa	1	11
Andorra	376	0
Angola	244	0
Anguilla	1	11
Antigua and Barbuda	1	11
Argentina	54	0
Armenia	374	0
Aruba	297	0
Australia	61	11
Austria	43	0
Azerbaijan	994	810
Bahamas	1	11
Bahrain	973	0
Bangladesh	880	0
Barbados	1	11
Belarus	375	810
Belgium	32	0

ISO 2 ALPHA	ISO 3 ALPHA	COUNTRY
AF	AFG	AFGHANISTAN
AG	ATG	ARGENTINA
ID	IDN	INDONESIA
IR	IRN	IRAN
IQ	IRQ	IRAQ
IE	IRL	IRELAND
IM	IMN	ISLE OF MAN
IL	ISR	ISRAEL
IT	ITA	ITALY
CI	CIV	CÔTE D'IVOIRE
JM	JAM	JAMAICA
JP	JPN	JAPAN
AR	ARG	ARMENIA
JE	JEY	JERSEY
JO	JOR	JORDAN
KZ	KAZ	KAZAKHSTAN
KE	KEN	KENYA
KI	KIR	KIRIBATI
KR	KOR	KOREA (Republic of [South] Korea)
KW	KWT	KUWAIT
KG	KGZ	KYRGYZSTAN
LA	LAO	LAO PEOPLE'S DEMOCRATIC REPUBLIC
AM	ARM	ARUBA
LV	LVA	LATVIA

POSTAL_CODE	CITY	STATE	ISO 2_CODE	ISO 3_CODE
210	Portsmouth	NH	US	USA
211	Portsmouth	NH	US	USA
212	Portsmouth	NH	US	USA
213	Portsmouth	NH	US	USA
214	Portsmouth	NH	US	USA
215	Portsmouth	NH	US	USA
501	Holtsville	NY	US	USA
544	Holtsville	NY	US	USA
601	Adjuntas	PR	US	USA
602	Aguada	PR	US	USA
603	Aguadilla	PR	US	USA
604	Aguadilla	PR	US	USA
605	Aguadilla	PR	US	USA
606	Maricao	PR	US	USA
607	Agua Buenas	PR	US	USA
609	Aibonito	PR	US	USA
610	Anasco	PR	US	USA
611	Angeles	PR	US	USA
612	Arecibo	PR	US	USA
613	Arecibo	PR	US	USA
614	Arecibo	PR	US	USA
615	Arroyo	PR	US	USA
616	Bajadero	PR	US	USA
617	Barceloneta	PR	US	USA
618	Barranquitas	PR	US	USA
622	Boqueron	PR	US	USA
623	Cabo Rojo	PR	US	USA
624	Penuelas	PR	US	USA
625	Caguas	PR	US	USA



Use master data to enrich your source data and also fill in missing values, key in completing address information....

Data Quality

Lastly, we briefly discussed the importance of master data and data quality processes such as data profiling and data cleansing.

DATE: 10/10/2012

DATA PROFILE REPORT

TBL NAME	COL NAME	NULLS	DUPLICATES	OUT OF RANGE	MAX VALUE	MIN VALUE	DISTINCT VALUES
State	State Key	0	0	0	1	50	51
	State Code	1	2	2	AL	??	49
	State Name	0	2	0	ALABAMA	WYOMING	50
	Country Code	2	0	1	US	ZZ	49

ROWS SAMPLED 51

Statistics Score	COL NAME	NULLS	DUPLICATES	Out Of Range	DISTINCT VALUES
State	State Key	0.00%	0.00%	0.00%	100.00%
	State Code	1.96%	3.92%	3.92%	96.08%
	State Name	0.00%	3.92%	0.00%	98.04%
	Country Code	3.92%	0.00%	1.96%	96.08%

DATE: 10/12/2012

DATA PROFILE REPORT

TBL NAME	COL NAME	NULLS	DUPLICATES	OUT OF RANGE	MAX VALUE	MIN VALUE	DISTINCT VALUES
State	State Key	0	0	0	1	50	50
	State Code	0	0	0	AL	WYOMING	50
	State Name	0	0	0	ALABAMA	WYOMING	50
	Country Code	0	0	0	US	US	50

ROWS SAMPLED 50

Statistics Score	COL NAME	NULLS	DUPLICATES	Out Of Range	DISTINCT VALUES
State	State Key	0.00%	0.00%	0.00%	100.00%
	State Code	0.00%	0.00%	0.00%	100.00%
	State Name	0.00%	0.00%	0.00%	100.00%
	Country Code	0.00%	0.00%	0.00%	100.00%

These architectures will deliver various reports and dashboards to report on data quality. Below is a mock up of a simple data quality scorecard:

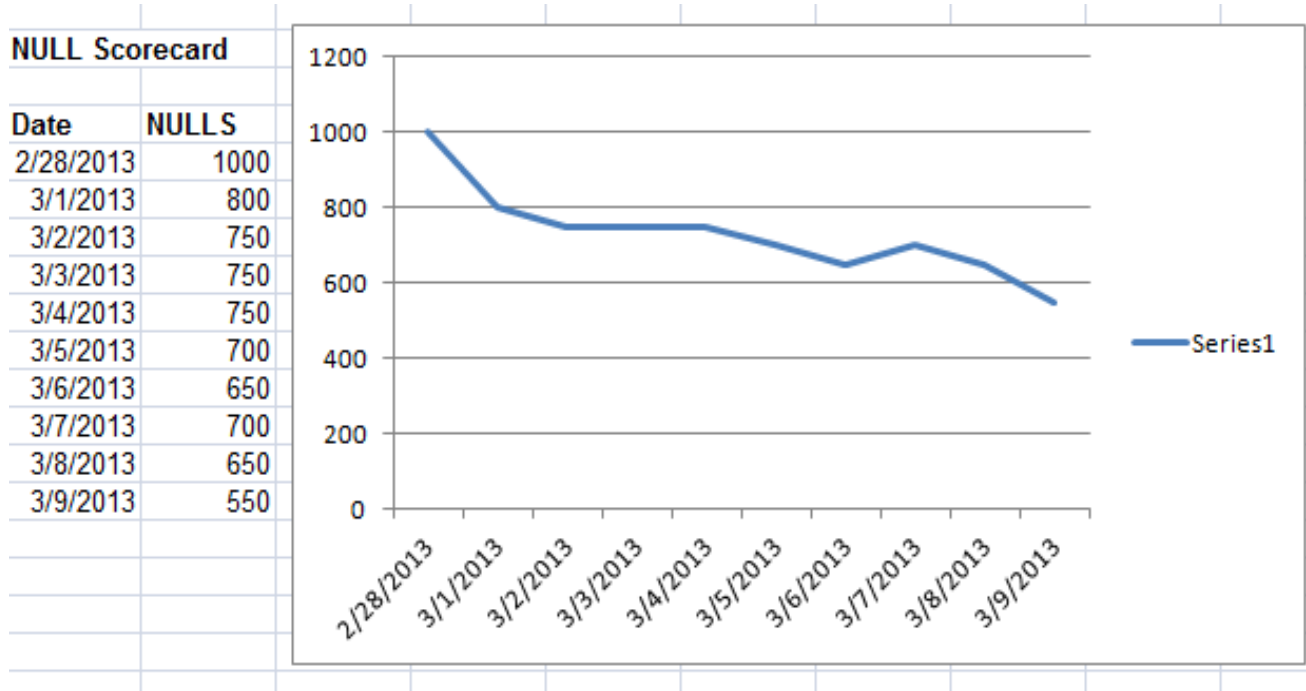
What to Check For....

Notice the basic checks:

- Checking for NULL values
- Checking for duplicate values
- Check maximum to minimum value range
- Check number of distinct values
- Check for out of bounds values
- Check for domain violations
- Check for missing primary-foreign key relationships

These checks provide the bare minimum information to deliver a profile of the quality of the data being examined. A before and after snapshot is taken to see if the data cleansing processes we implement are working.

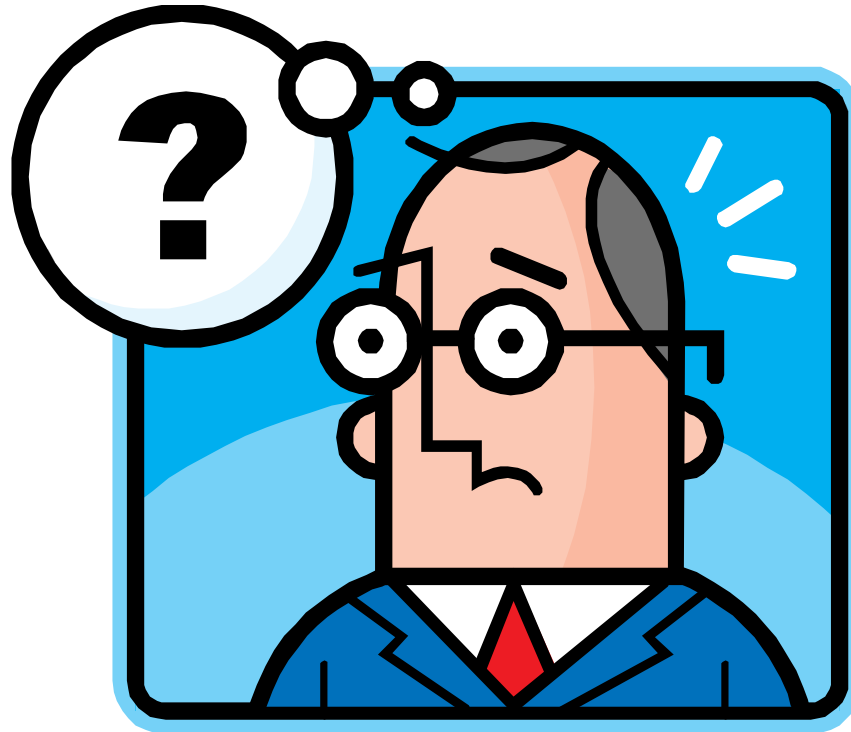
Additional scorecards....



This is an excellent method to present data stewards with the levels of data quality so they can identify issues, recommend processes for cleansing and monitor results.

This is also a valuable tool in the schema integration processes as it allows us to check the quality of the incoming data as we load it into the new ODS

Thank you for your attention!



Here is a promo code for 30% off Connecting The Data: **ConnectingData30**

You can order the book from the Technics Publications website at this link:

<http://www.technicspub.com/product.sc;jsessionid=629EA516EFC906CE6EDEC59BA39F46E.qscstrfrnt01?productId=52&categoryId=1>

And then enter the coupon code upon checkout.