E T L

Day-3 – Extraction

Cyrus Lentin

What Is Required?

- Crucial To Know How To Extract The Data From The Source System
- Each Source Has Distinctive Characteristics And Need To Be Managed Accordingly
- Integration With Different Systems
  - Database Management System
  - Operating Systems
  - Hardware
  - Communication Protocols
- It Is Important To Logically Map The Data From The Source Defining Different Fields

Extracting Data – Considerations

- Understanding Data - Logical Data Map
- Extracting Data From Disparate Platforms
- Extracting Changed Data

Understanding Data – Logical Data Map

- Logical Data Plan Before Physical Activity
- Ensure The Following Steps Are Achieved Before You Start Any Physical ETL Development:
  - Have A Plan
  - Identify Data Source Candidates
  - Analyze Source Systems With A Data-profiling Tool
  - Receive Walk-though Of Data Lineage And Business Rules
  - Receive Walk-through Of Data Warehouse Data Model
  - Validate Calculations And Formulas

Components Of Logical Data Map

- Table Name
- Column Name
- Table Type (Fact, Dimension, Sub Dimension, Supporting Etc)
- Slow Changing Dimension Type (Applicable For Dimension Tables)
- Source Database To Get This Information From
- Source Table
- Source Column Name
- Transformation Required (If Any)

Logical Data Map Report

Target					Source				Transformation
Table Name	Column Name	Data Type	Table Type	SCD Type	Database Name	Table Name	Column Name	Data Type	
EMPLOYEE_DIM	EMPLOYEE_KEY	NUMBER	Dimension	1				NUMBER	Surrogate key.
EMPLOYEE_DIM	EMPLOYEE_ID	NUMBER	Dimension	1	HR_SYS	EMPLOYEES	EMPLOYEE_ID	NUMBER	Natural Key for employee in HR system
EMPLOYEE_DIM	BIRTH_COUNTRY_NAME	VARCHAR2(75)	Dimension	1	HR_SYS	COUNTRIES	NAME	VARCHAR2(75)	select c.name from employees e, states s, countries c where e.state_id = s.state_id and s.country_id = c.country
EMPLOYEE_DIM	BIRTH_STATE	VARCHAR2(75)	Dimension	1	HR_SYS	STATES	DESCRIPTION	VARCHAR2(255)	select s.description from employees e, states s where e.state_id = s.state_id
EMPLOYEE_DIM	DISPLAY_NAME	VARCHAR2(75)	Dimension	1	HR_SYS	EMPLOYEES	FIRST_NAME	VARCHAR2(75)	select initcap(salutation) || ' '||initcap(first_name) || ' '|| initcap(last_name) from employee
EMPLOYEE_DIM	BIRTH_DATE	DATE	Dimension	1	HR_SYS	EMPLOYEES	DOB	DATE	trunc(DOB)
EMPLOYEE_DIM	SALUTATION	VARCHAR2(12)	Dimension	1	HR_SYS	EMPLOYEES	SALUTATION	VARCHAR2(12)	initcap(salutation)
EMPLOYEE_DIM	FIRST_NAME	VARCHAR2(30)	Dimension	1	HR_SYS	EMPLOYEES	FIRST_NAME	VARCHAR2(30)	initcap(first_name)
EMPLOYEE_DIM	LAST_NAME	VARCHAR2(30)	Dimension	1	HR_SYS	EMPLOYEES	LAST_NAME	VARCHAR2(30)	initcap(last_name)
EMPLOYEE_DIM	MARITAL_STATUS	VARCHAR2(12)	Dimension	2	HR_SYS	MARITAL_STATUS	DESCRIPTION	VARCHAR2(12)	select nvl(m.name,'Unknown') from employee e marital_status m where e.marital_status_id = m.marital_status_id
EMPLOYEE_DIM	DIVERSITY_CATEGORY	VARCHAR2(30)	Dimension	1	HR_SYS	EMPLOYEES	EEO_CLASS	VARCHAR2(30)	decode(eeo_class,null, 'Not Stated', decode(eeo_class,'N', 'Not Stated',eeo_class))
EMPLOYEE_DIM	GENDER	VARCHAR2(12)	Dimension	1	HR_SYS	EMPLOYEES	SEX	VARCHAR2(12)	nvl(sex, 'Unknown')
EMPLOYEE_DIM	EMPLOYEE_STATUS	VARCHAR2(24)	Dimension	1	HR_SYS	EMPLOYEES	STATUS	VARCHAR2(24)	select es.name from employee e employee_status es where e.employee_status_id = m.employee_status_id
									select p.code from employees e, positions p where

Important Factors Of Loading

- Important Factor To Be Considered When Loading The Dimension Tables
- Structure Of The Dimension Table Cannot Tell What The Strategy Is
- Columns Have Historic Relevance And The Strategy Required For Capturing This History Should Be Known In Advance
- Changing The SCD After The Design Should Be Managed Well Through A Change Management Process
- SCDs Data Changes Slowly, Rather Than Changing On Regular Schedule
- Example:
  - City
  - Country

Slowly Changing Data (SCD)

- Type 0
  Passive Values remain same for ever
- Type 1
  Allows new data to overwrite old data So not required to track the history
- Type 2
  Tracks historical data by creating multiple records for a given natural key in the dimensional tables with separate surrogate keys and/or different version numbers
- Type 3
  Tracks changes using separate columns and preserves limited history
- Type 4
  Maintains older data in separate history tables
- Type 5
  Also called hybrid It is a combination of 2 or more of the above methods

Building Logical Data Map

- Logical Data Mapping Not Possible Until The Source Systems Have Been Identified And Analyzed
- The Analysis Of The Source System Is Usually Broken Into Two Major Phases:
  - Data Discovery Phase
    - Collecting and Documenting Source Systems – Entity Relationship Diagram
    - Keeping Track of the Source Systems – Source System Tracking Report
  - Anomaly Detection Phase
    - Data Content Analysis
    - Collecting Business Rules
    - Collecting Heterogenous Data

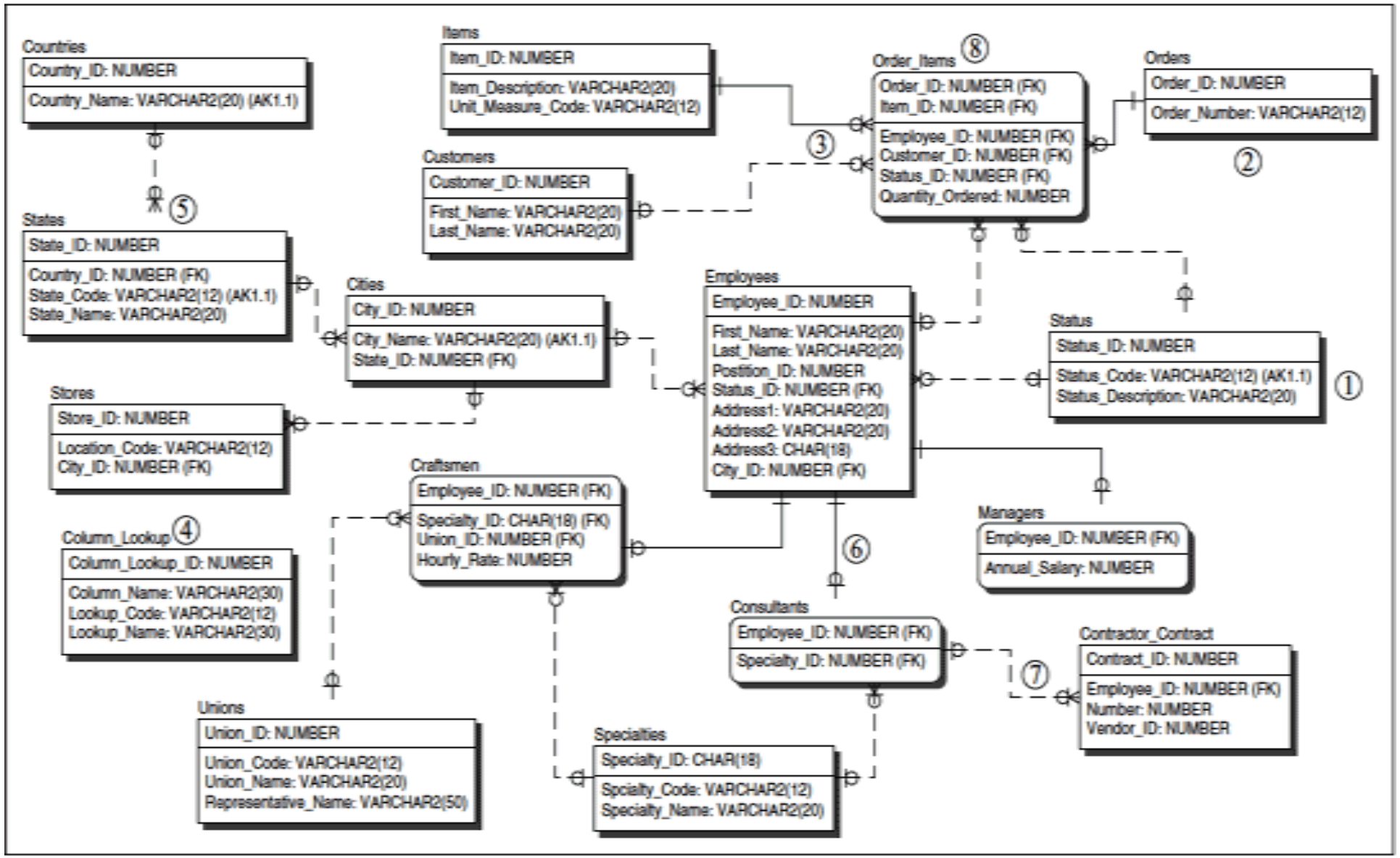
Entity Relationship Diagram

- Unique Identifiers And Natural Keys
- Data Types
- Relationships Between Tables
- Discrete Relationships
- Cardinality Of Relationships And Columns
  - One-to-one
  - One-to-many
  - Many-to-many

Important

- Referential Key Integrity
- Data Type Issues Between Source & ETL

Entity Relationship Diagram



Source System Tracking Report

- Subject Area
- Interface Name
- Business Name
- Priority
- Department / Business Use
- Business Owner
- Technical Owner
- DBMS
- Platform
- Daily Users Count
- Db Size
- Transactions Per Day
- Comments

Source System Tracking Report

Subject Area	Interface Name	Business Name	Priority	Department/ Business Use	Business Owner	Technical Owner	DBMS	Platform	Daily User Count	DB Size (GB)	Transactions/ Day	comments
Human Resources	PeopleSoft	HR Information	1	Human Resource Mgmt	Daniel Burnis	Margaret Karlin	Oracle	AIX on RS/6000	850	3	8,000	Everything revolves around HR
Finance	Oracle Financials	Oracle Financials	2	General Accounting	Mabel Karr	Edmund Bykel	Oracle	AIX on S80	400	60	20,000	Driven sponsor ; Department likes slicing & dicing data
Materials Management	WMS	Warehouse Management System	3	Manufacturing Planning & Control	Annabel Giutu	Christina Hayim	Oracle	AIX on S80	350	200	5,000	Initiative to cut inventory of slow movers; Need analytical support.
Operations	BOMM	Bill of Materials Management	4	Production Planning	Lucy Kard	George Cimentbiis	Oracle	AIX on RS/6000	60	8	TBD	Need analysis of "kits", Want to build kits in-house
Marketing	IMCM	Internet Marketing & Campaign Management	5	Marketing	Elizabeth Impzar	Brian Bridalurg	SQL Server	W2K on Compaq	50	350	Varies	Needs Data Mining Capabilities
Human Resources	Positions Website	Positions Website	5	Human Resource Mgmt	Florence Iyzar	Andrew Aoenthal	SQL Server	W2k on Compaq	50	1	10,000	Manager is new, eager for DW
Purchasing	PAS	Purchasing & Acquision System	6	Purchasing Department	Guy Croendil	Sybil Mai	SQL Server	NT on Compaq	75	4	1,500	
Customer Service	CSS	Customer Service System	7	Customer Care	Bernard Beais	Arthur Aficcio	Notes/ Domino	NT on Compaq	275	6	1,500	Plans to move application to Oracle next year.
Operations	QCS	Quality Control System	8	Quality Control	Thomas Fryar	Elias Caye	SQL Server	NT on Compaq	450	12	500	Needs failure ratio analytics by product, by vendor, by year
Sales	SFA	Sales Force Automation	9	Sales	Anthony Aran	Emma Inalonn	DB2	AS/400	1200	10	2,500	Politically challenging; Interim reporting system in place

Data Content Analysis

- Referential Key Integrity
- Data Type Issues Between Source & ETL
- NULL Values
- Date Values

Business Rules

- Status
- Flags
- Labels
- Classification

Important

- Undocumented

Integrating Heterogeneous Data Sources

- Identify the source systems
- Understand the source systems (data profiling)
- Create record matching logic
- Establish survivorship rules
- Establish non-key attribute business rules
- Load conformed dimension

Extracting Data from Disparate Platforms

- ODBC
- Mainframe
- Flat Files
- XML
- Web Logs
- ERP Systems

O D B C

- Open Database Connectivity (ODBC)
- Was created to enable users to access databases from their applications retrospective of the OS
- The original intention for ODBC was to make applications portable, meaning that if an application's underlying database changed—say from DB2 to Oracle—the application layer did not need to be recoded and compiled to accommodate the change
- Instead, you simply change the ODBC driver, which is transparent to the application You can obtain ODBC drivers for practically every DBMS in existence on virtually any platform You can also use ODBC to access flat files

Advantage

- Flexibility
- Interoperability

Drawback

- Performance
- DBMS Based Commands

Mainframe

- In Many Large Companies, Much Of The Day-to-day Business Data Is Processed And Stored On Mainframe Systems (And Certain Minicomputer Systems, Such As The IBM AS/400) And Integrating Data From These Systems Into The Data Warehouse Involves Some Unique Challenges
- Mainframes Have A Special Architecture Emphasizing Peripheral Channels That Process All Input/Output, Leaving The CPU Dedicated To Processing Only Data, Such As Calculating Formulas And Balances
- There Are Several Characteristics Of Mainframe Systems That The ETL Team Must Be Familiar With And Develop Techniques To Handle:
  - COBOL Datasets
  - EBCDIC Character Sets
  - Numeric Data
  - Packed Decimal Fields
  - Multiple Record Types
  - Variable Record Lengths

Flat Files

- Flat files are the mainstay of any data-staging application
- Flat files are utilized by the ETL process for at least three reasons:
  - Delivery Of Source Data
  - Working / Staging Tables
  - Preparation For Bulk Load
- Flat files essentially come in two flavors:
  - Fixed length
  - Delimited

Flat Files – Pre Validation

Fixed Length Flat Files – Pre Validation

- Validate That The Positions Of The Data In The File Are Accurate
- Check For The Positions Is To Test Any Date (Or Time) Field To Make Sure It Is A Valid Date
- If The Positions Are Shifted, The Date Field Most Likely Contains Alpha Characters Or Illogical Numbers
- Other Class / Category / Status / Flags Fields With Can Be Tested In The Same Way
- Positional Flat Files Are Often Indicated On The File System By A TXT Extension
- Positional Flat Files Can Have Virtually Any File Extension—or None At All

Delimited Files – Pre Validation

- Validate That The Delimiters In The File Are In Order
- Check For The Delimiters To Test Any Date (Or Time) Field To Make Sure It Is A Valid Date
- If The Delimiters Are Shifted, The Date Field Most Likely Contains Alpha Characters Or Illogical Numbers
- Other Class / Category / Status / Flags Fields With Can Be Tested In The Same Way
- Delimited Files Are Often Indicated On The File System By A CSV Extension
- Delimited Files Also Can Have Virtually Any File Extension—or None At All

Must Have Explicit Validation Tests Written By The ETL Team And Embedded In The ETL Routines

X M L

- Extensible Markup Language (XML) Is Slowly But Surely Becoming The Standard For Sharing Data
- XML Has Emerged To Become A Universal Language For Exchanging Data Between Enterprises
- If Your Data Warehouse Includes Data That Comes From External Sources—those From Outside Of Your Enterprise—odds Are That Those Sources Will Be Provided In XML
- XML Has Two Important Elements:
  - Metadata
  - Data
- Xml Considerations
  - Character Sets
    - Character Sets Are Groups Of Unique Symbols Used For Displaying And Printing Output
    - Default Character Set For Most Relational Database Management Systems Is Iso8859-15 (Latin 9)
    - XML Supports The Utf-8 Character Set UTF-8 Supports Most Of The Languages And Alphabets
  - XML Meta Data
    - XML Schema
    - Namespace

      Always Use XML Reader Ensure ETL Tool Can Natively Process XML And XML Schemas

Web Logs

- Web logs are line records of every request to & response from the Web server
- Web logs are important because they reveal the user traffic & many other details on the Web site
- Fortunately, the format of the text-based log is standardized by W3C

W3C Common Format

- Date
- client-ip
- server-ip
- method
- url
- query
- protocol
- sc-status
- user-agent
- referrer

W3C Extended Formats

- log-server
- username
- server-port
- time-taken

ERP Systems

- ERP systems were created to solve one of the issues that data warehouses face today—integration of heterogeneous data
- ERP systems are designed to be an integrated enterprise solution that enables every major entity of the enterprise, such as sales, accounting, human resources, inventory, and production control, to be on the same platform, database, and application framework
- ERP systems are extremely complex and not easily implemented They take months or years to customize so they contain the exact functionality to meet all of the requirements to run a particular business
- ERP systems are notoriously large, and because they are really a framework and not an application, their data models are comprehensive, often containing thousands of tables
- Moreover, because of their flexibility, the data models that support ERP processing are incredibly difficult to navigate
- Inspite of the effort to be an all-inclusive solution, it's very rare to see an entire enterprise use only an ERP system to run a company
- The more popular ERP systems are SAP, PeopleSoft, Oracle, Baan, and JDEdwards

Many of the major ETL vendors now offer ERP adapters to communicate with the popular ERP systems

Extracting Changed Data

- The initial load of ETL is easy
- Once that load is complete, the ability to capture data changes in the source system becomes priority
- Capturing data changes is very complicated task
- You must plan your strategy to capture incremental changes to the source data at the onset
- Steps Involved
  - Detecting Changes
  - Extraction Tips
  - Detecting Deleted or Overwritten Fact Records at the Source

Detecting Changes

- Using Audit Columns
- Database Log Scraping Or Sniffing
- Timed Extracts
- Process Of Elimination
- Initial And Incremental Loads

Header

- Bullets

Extraction Tips

- Use DISTINCT Sparingly
- Use SET Operators Sparingly
- Use HINT / EXPLAIN As Necessary
- Avoid Not
- Avoid Functions In Your Where Clause

Detecting Deleted Or Overwritten Fact Records At The Source

- Measurement (fact) records deleted or overwritten from source systems can pose a very difficult challenge for the data warehouse if no notification of the deletion or overwrite occurs.
- Since it is usually infeasible to repeatedly re-extract old transaction records, looking for these omissions and alterations, the best we can offer are the following procedures:
  - Negotiate with the source system owners, if possible, explicit notification of all deleted or overwritten measurement records.
  - In cases of deleted or modified fact records, rather than just performing a deletion or update in the data warehouse, a new record is inserted that implements the change in the fact by canceling or negating the originally posted value. In many applications, this will sum the reported fact to the correct quantity (if it is additive) as well as provide a kind of audit trail that the correction occurred.
  - Periodically check historical totals of measurements from the source system to alert the ETL staff that something has changed. When a change is detected, drill down as far as possible to isolate the change.
- In these cases, it may also be convenient to carry an extra administrative time stamp that identifies when the database actions took place.

Thank you!

Contact:

Cyrus Lentin
cyrus@lentinscoin
+91-98200-94236