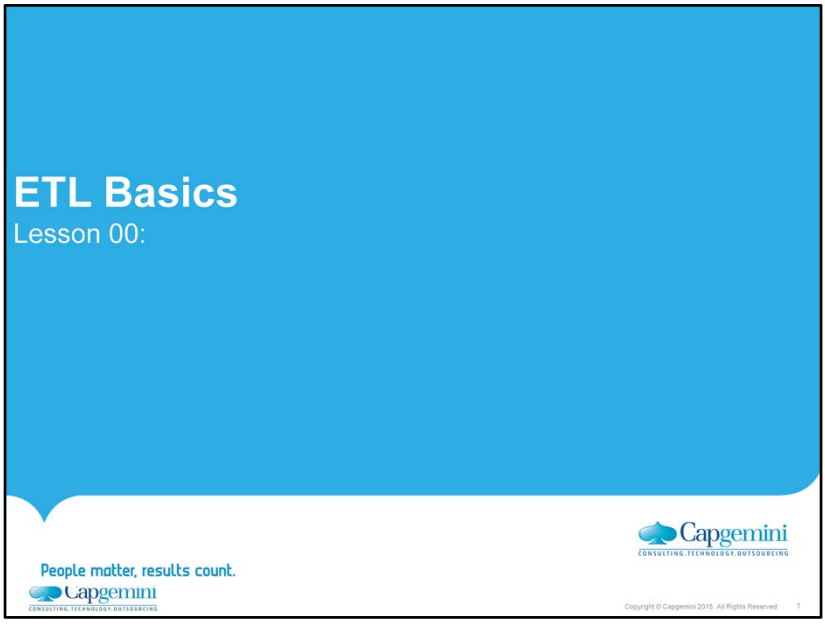


Document History

Date	Course Version No.	Software Version No.	Developer / SME	Reviewer(s)	Approver	Change Record Remarks
June 2011	1	NA	Vandana Mistry			Content Creation
July 2016	1.1	NA	Swati Rao	Rajita Dhumal	Mahima Sharma	Material Revamp as per Integrated ToC for I & D LoT

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

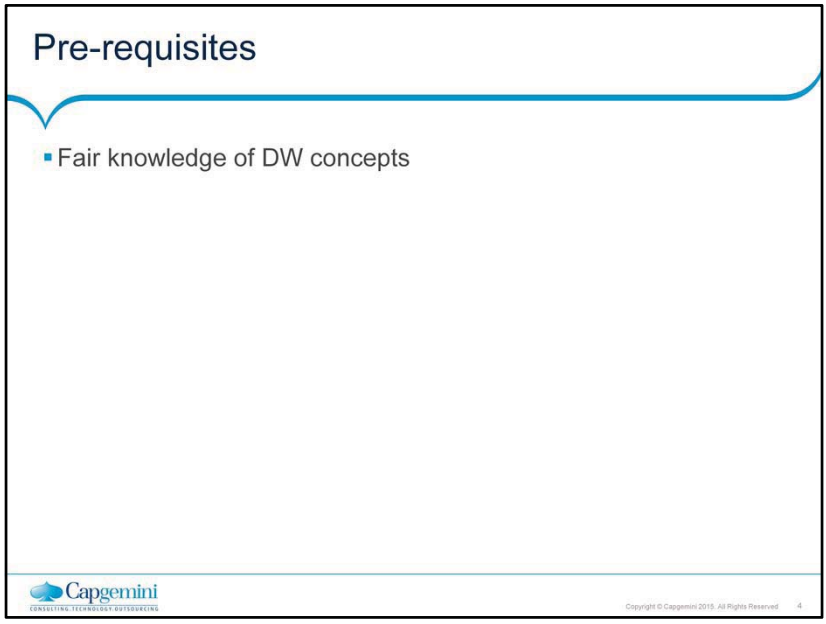
 2

Course Goals and Non Goals

- Course Goals
  - At the end of this program, participants gain an understanding of basic concepts in ETL.
- Course Non Goals
  - Implementation of ETL tools.

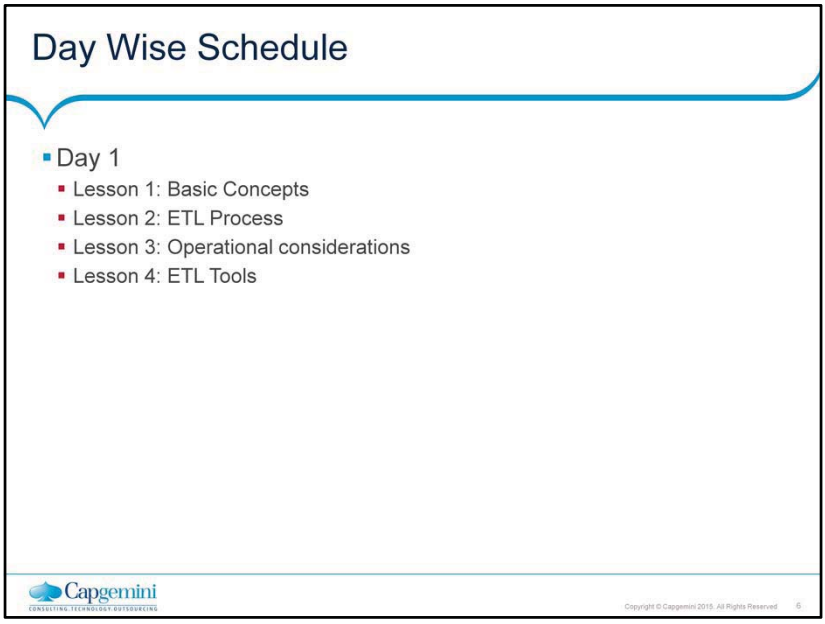
Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

 3


Pre-requisites
Fair knowledge of DW concepts
Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING
Copyright © Capgemini 2015. All Rights Reserved
4

Intended Audience

- Software Engineers and Senior Software Engineers


Day Wise Schedule
Day 1
Lesson 1: Basic Concepts
Lesson 2: ETL Process
Lesson 3: Operational considerations
Lesson 4: ETL Tools
Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING
Copyright © Capgemini 2015. All Rights Reserved
6

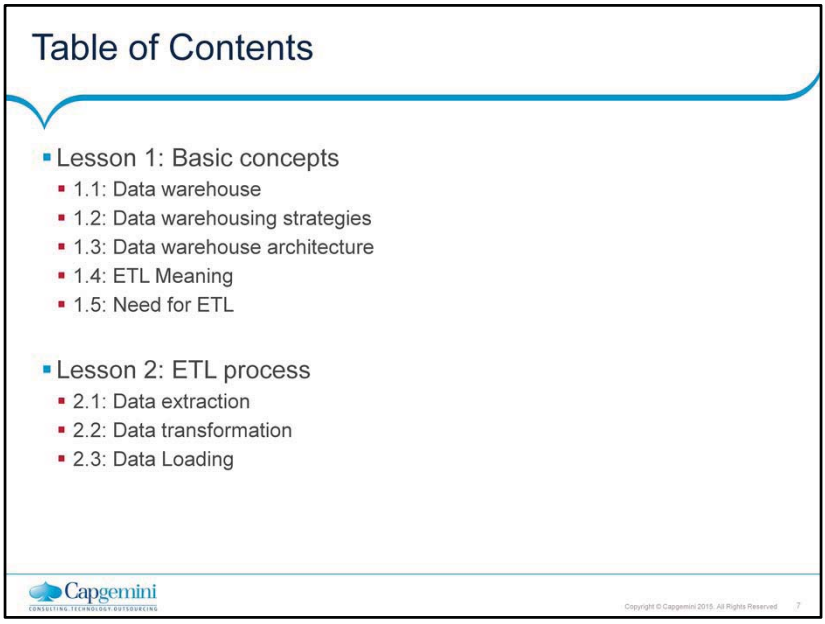


Table of Contents
Lesson 1: Basic concepts
1.1: Data warehouse
1.2: Data warehousing strategies
1.3: Data warehouse architecture
1.4: ETL Meaning
1.5: Need for ETL
Lesson 2: ETL process
2.1: Data extraction
2.2: Data transformation
2.3: Data Loading
Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING
Copyright © Capgemini 2015. All Rights Reserved
7

Table of Contents

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

References

- Student material:
  - Class Book (presentation slides with notes)
- Book:
  - The Data Warehousing ETL Toolkit – Ralph Kimball
- Web-site:
  - http://www.datawarehouse.org
  - http://etl-tools.info/

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

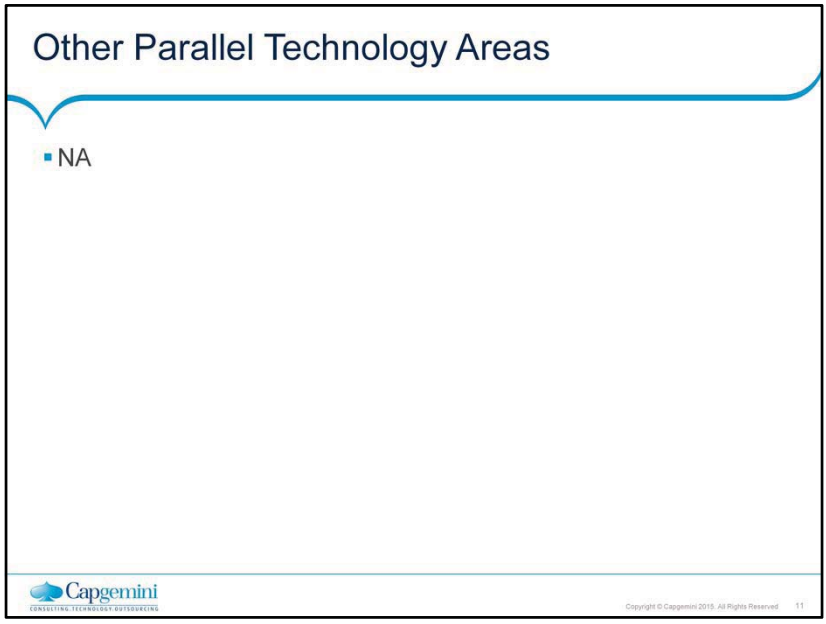
 9

Next Step Courses (if applicable)

- BI related tool training

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

 10


Other Parallel Technology Areas
NA
Capgemini
CONSULTING. TECHNOLOGY. OUTSOURCING
Copyright © Capgemini 2015. All Rights Reserved
11

ETL Basics

Lesson 1: Basic Concepts

Lesson Objectives

- On completion of this lesson on ETL basics, you will be able to:
  - Understand Data warehousing strategies and architecture
  - Know the meaning and need of ETL

Capgemini
CONSULTING. TECHNOLOGY. OUTSOURCING

 2

Datawarehouse

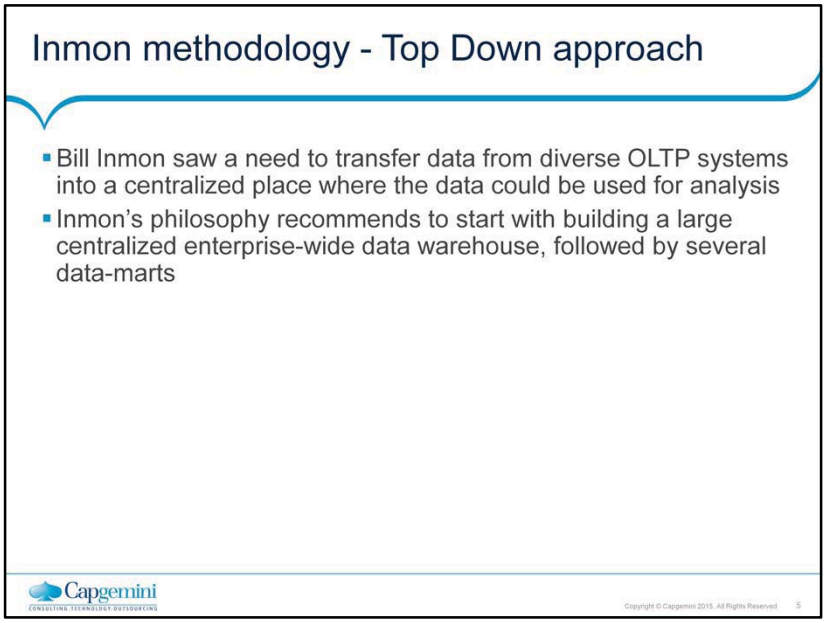
- A single, complete and consistent store of data obtained from a variety of different sources made available to end users in a format that they can understand and use in a business context.

Datawarehousing Strategies

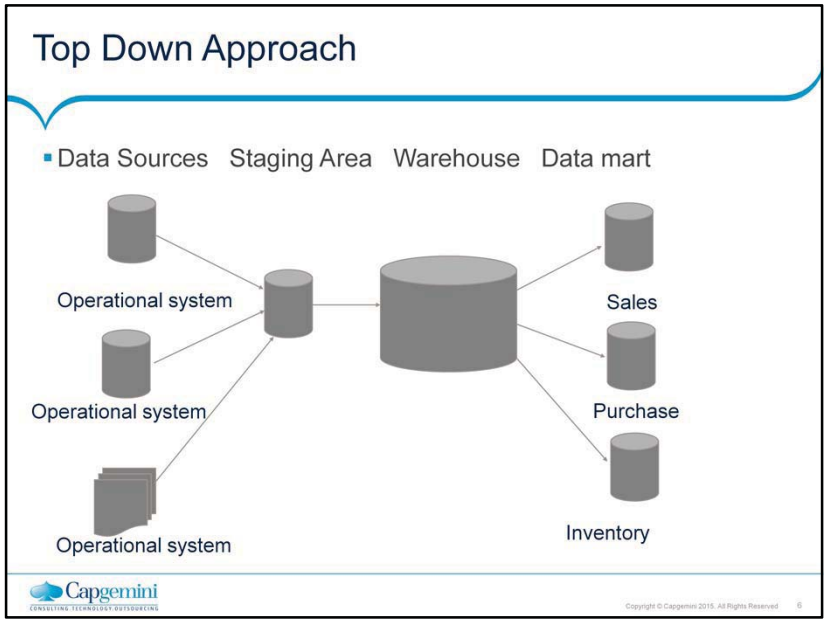
- Enterprise-wide warehouse, top down, the Inmon methodology
- Data mart, bottom up, the Kimball methodology
- When properly executed, both result in an enterprise-wide data warehouse

Capgemini
CONSULTING. TECHNOLOGY. OUTSOURCING

 4



The data marts are treated as sub sets of the data warehouse. Each data mart is built for an individual department and is optimized for analysis needs of the particular department for which it is created. The data flow in the top down OLAP environment begins with data extraction from the operational data sources. This data is loaded into the staging area and validated and consolidated. This data from the Staging area is then loaded in to the datawarehouse.



The data flow in the top down OLAP environment begins with data extraction from the operational data sources. This data is loaded into the staging area and validated and consolidated for ensuring a level of accuracy and then transferred to the Operational Data Store. (ODS). The ODS stage is sometimes skipped if it is a replication of the operational databases. Data is also loaded into the Data warehouse in a parallel process to avoid extracting it from the ODS.

Detailed data is regularly extracted from the ODS and temporarily hosted in the staging area for aggregation, summarization and then extracted and loaded into the Data warehouse. The need to have an ODS is determined by the needs of the business. If there is a need for detailed data in the Data warehouse then, the existence of an ODS is considered justified. Else organizations may do away with the ODS altogether.

Once the Data warehouse aggregation and summarization processes are complete, the data mart refresh cycles will extract the data from the Data warehouse into the staging area and perform a new set of transformations on them. This will help organize the data in particular structures required by data marts. Then the data marts can be loaded with the data and the OLAP environment becomes available to the users.

The data in a data warehouse is time variant in nature as it contains historical data. Inmon proposes a top-down model approach to create a centralized Enterprise Data Warehouse using traditional database

modeling techniques (ER Model), where the data is stored in 3NF. The data warehouse acts as data source for the new data marts

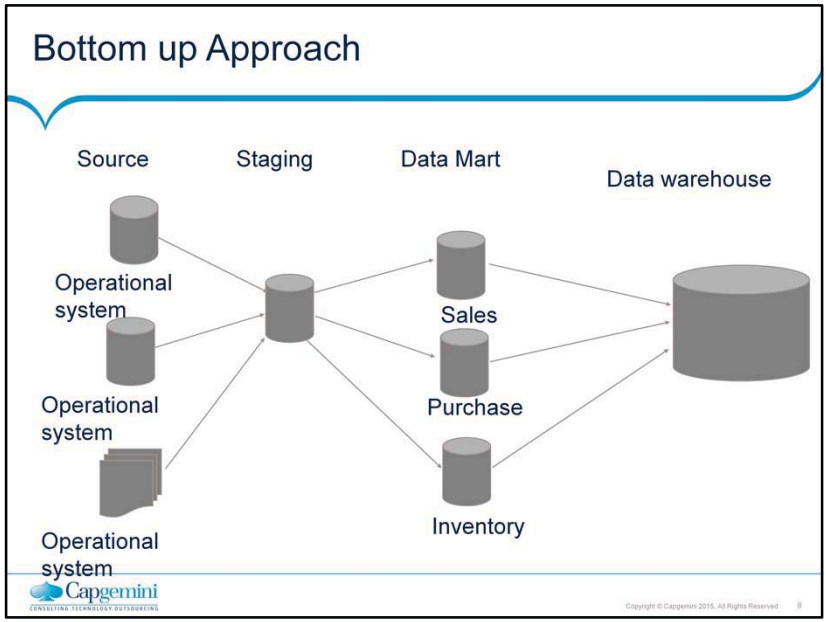
Kimball methodology – Bottom Up approach

- Kimball's philosophy recommends to start with building several data marts that serve the analytical needs of departments, followed by "virtually" integrating these data marts.

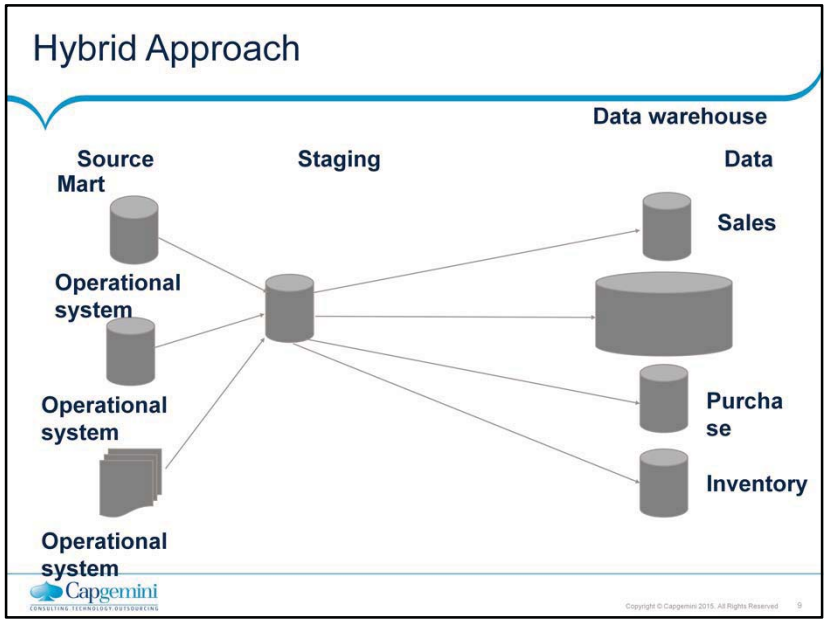
Capgemini
CONSULTING. TECHNOLOGY. OUTSOURCING

 7

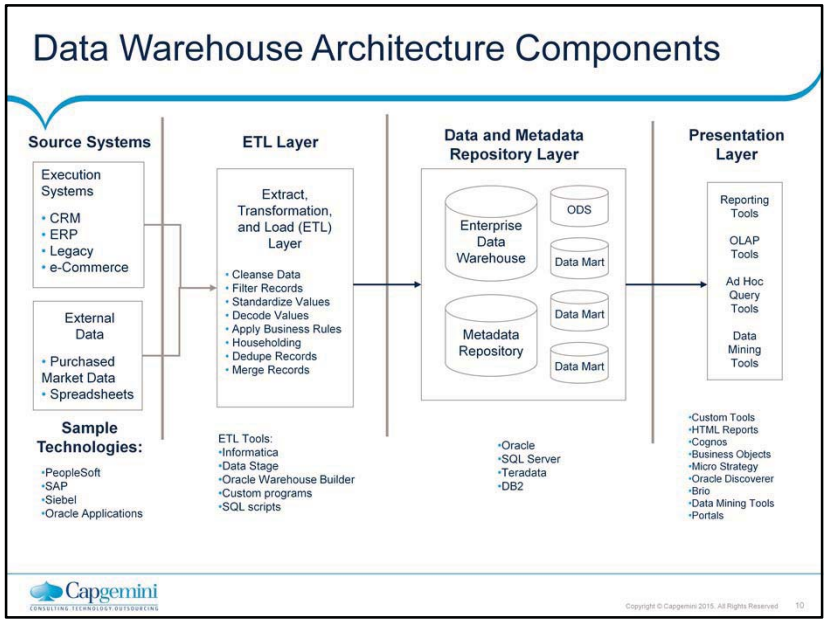
.



The bottom-up approach reverses the positions of the Data warehouse and the Data marts. Data marts are directly loaded with the data from the operational systems through the staging area. The ODS may or may not exist depending on the business requirements.
The data flow in the bottom up approach starts with extraction of data from operational databases into the staging area where it is processed and consolidated and then loaded into the ODS. The data in the ODS is appended to or replaced by the fresh data being loaded. After the ODS is refreshed the current data is once again extracted into the staging area and processed to fit into the Data mart structure. The data from the Data Mart, then is extracted to the staging area aggregated, summarized and so on and loaded into the Data Warehouse and made available to the end user for analysis.



The bottom-up approach reverses the positions of the Data warehouse and the Data marts. Data marts are directly loaded with the data from the operational systems through the staging area. The ODS may or may not exist depending on the business requirements.
The data flow in the bottom up approach starts with extraction of data from operational databases into the staging area where it is processed and consolidated and then loaded into the ODS. The data in the ODS is appended to or replaced by the fresh data being loaded. After the ODS is refreshed the current data is once again extracted into the staging area and processed to fit into the Data mart structure. The data from the Data Mart, then is extracted to the staging area aggregated, summarized and so on and loaded into the Data Warehouse and made available to the end user for analysis.


Data Warehouse Architecture Components
Source Systems
Execution Systems
• CRM
• ERP
• Legacy
• e-Commerce
External Data
• Purchased Market Data
• Spreadsheets
Sample Technologies:
•PeopleSoft
•SAP
•Siebel
•Oracle Applications
ETL Layer
Extract, Transformation, and Load (ETL) Layer
• Cleanse Data
• Filter Records
• Standardize Values
• Decode Values
• Apply Business Rules
• Householding
• Dedupe Records
• Merge Records
ETL Tools:
•Informatica
•Data Stage
•Oracle Warehouse Builder
•Custom programs
•SQL scripts
Data and Metadata Repository Layer
Enterprise Data Warehouse
ODS
Data Mart
Data Mart
Metadata Repository
Data Mart
•Oracle
•SQL Server
•Teradata
•DB2
Presentation Layer
Reporting Tools
OLAP Tools
Ad Hoc Query Tools
Data Mining Tools
•Custom Tools
•HTML Reports
•Cognos
•Business Objects
•Micro Strategy
•Oracle Discoverer
•Brio
•Data Mining Tools
•Portals
Capgemini
CONSULTING. TECHNOLOGY. OUTSOURCING
Copyright © Capgemini 2015. All Rights Reserved 10

What is ETL?

- ETL stands for Extract Transform & Load
- The process of updating the data warehouse
- ETL is the automated and auditable data acquisition process from source system that involves one or more sub processes of data extraction, data transportation, data transformation, data consolidation, data integration, data loading and data cleaning.

Capgemini
CONSULTING. TECHNOLOGY. OUTSOURCING

 11

Need for ETL

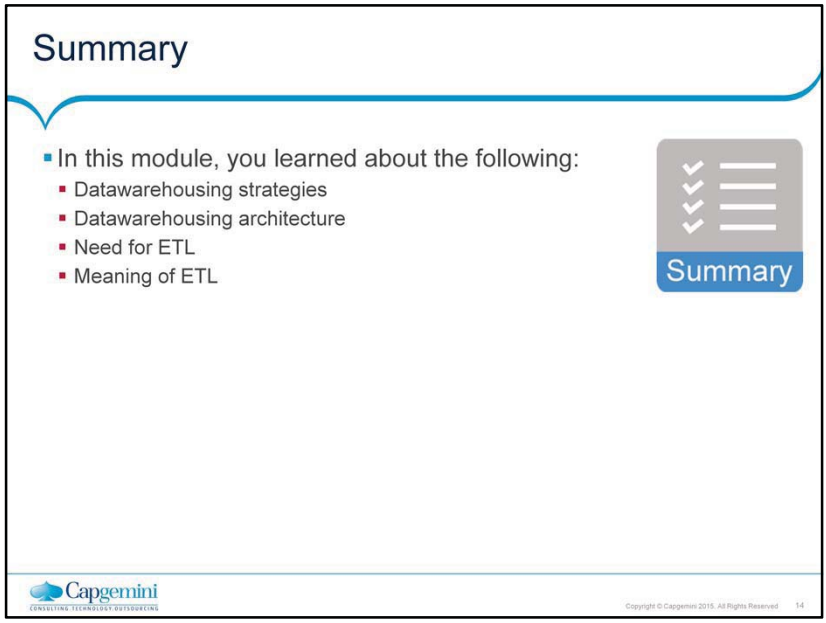
- The process of ETL is required so that data from different heterogeneous sources can be combined and brought into one common source.
- The Advantage of having the process of ETL is that, as data from different sources can be brought together, highly complex and user friendly reports can be generated for decision making

Need for ETL

- Data stored in different formats in different types of databases
- Some data sources might be archives while others may be active operational systems
- Data extraction and cleansing - time-consuming and difficult
  Aggregation of data

Capgemini
CONSULTING. TECHNOLOGY. OUTSOURCING

 13



Add the notes here.

ETL Basics

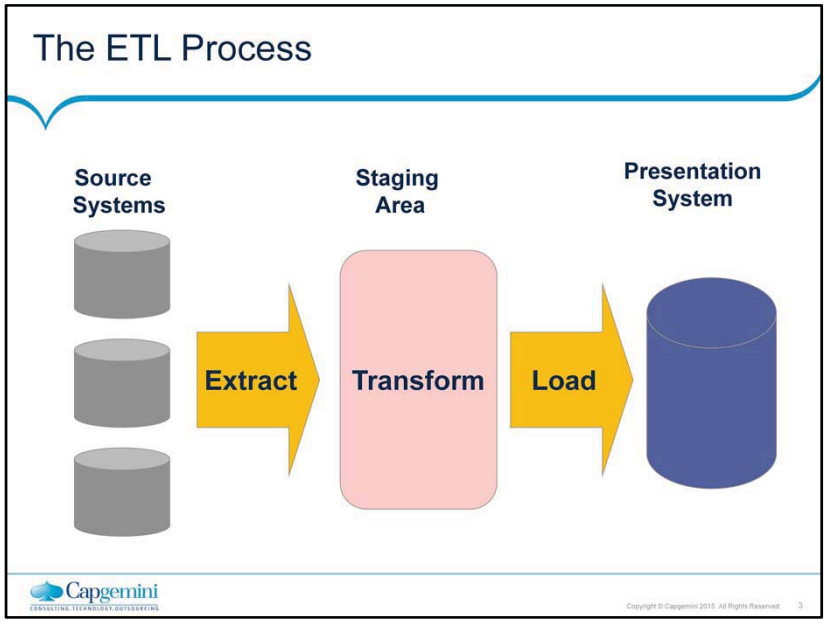
Lesson 2: ETL Process

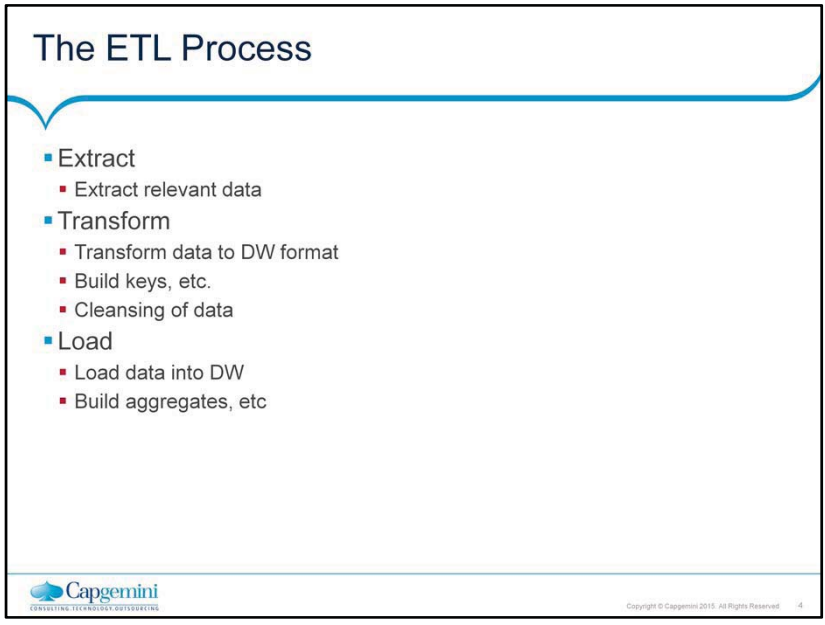
Lesson Objectives

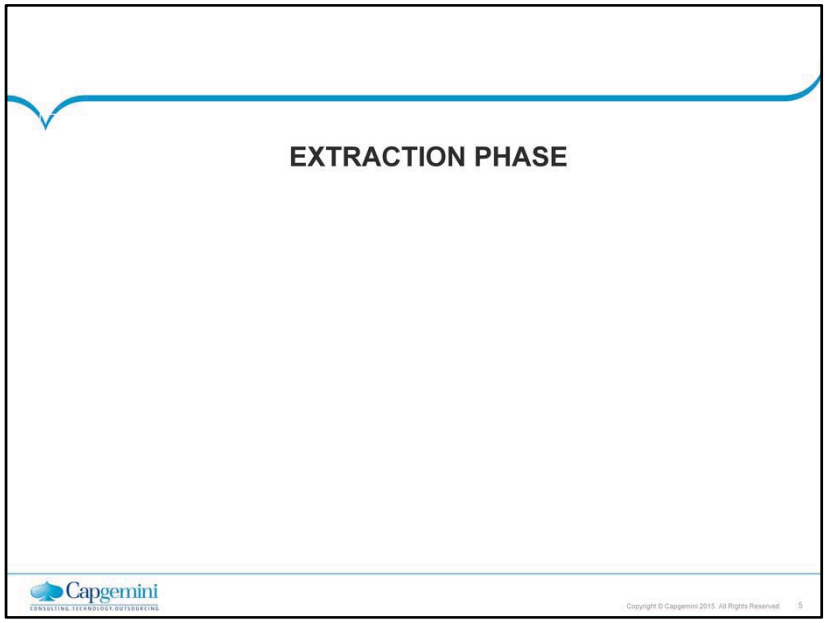
- On completion of this lesson on Data Modeling, you will be able to understand:
  - The ETL process
  - The steps in Data Cleansing

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

 2

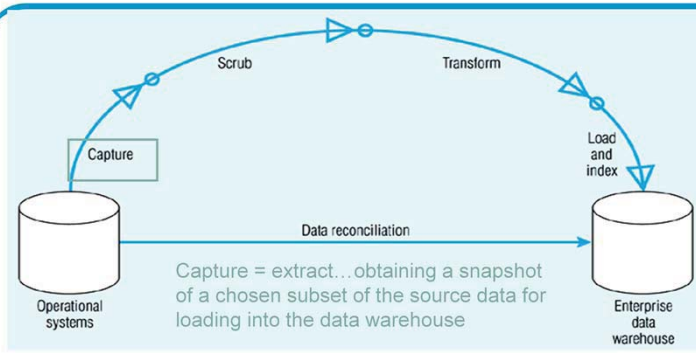

The ETL Process
Source Systems
Staging Area
Presentation System
Extract
Transform
Load
Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING
Copyright © Capgemini 2015. All Rights Reserved
3


The ETL Process
Extract
Extract relevant data
Transform
Transform data to DW format
Build keys, etc.
Cleansing of data
Load
Load data into DW
Build aggregates, etc
Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING
Copyright © Capgemini 2015. All Rights Reserved. 4



Now Let' s go through that how Transforming data will take place in the Data Warehousing environment

ETL – DATA CAPTURE



Static extract = capturing a snapshot of the source data at a point in time

Incremental extract = capturing changes that have occurred since the last static extract

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

 6

Change Data Capture

- Data warehousing involves the extraction and transportation of data from one or more databases into a target system or systems for analysis.
- But this involves the extraction and transportation of huge volumes of data and is very expensive in both resources and time.
- The ability to capture only the changed source data and to move it from a source to a target system(s) in real time is known as Change Data Capture (CDC).

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

 7

Change Data Capture

- CDC helps identify the data in the source system that has changed since the last extraction.
- Set of software design patterns used to determine the data that has changed in a database.

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

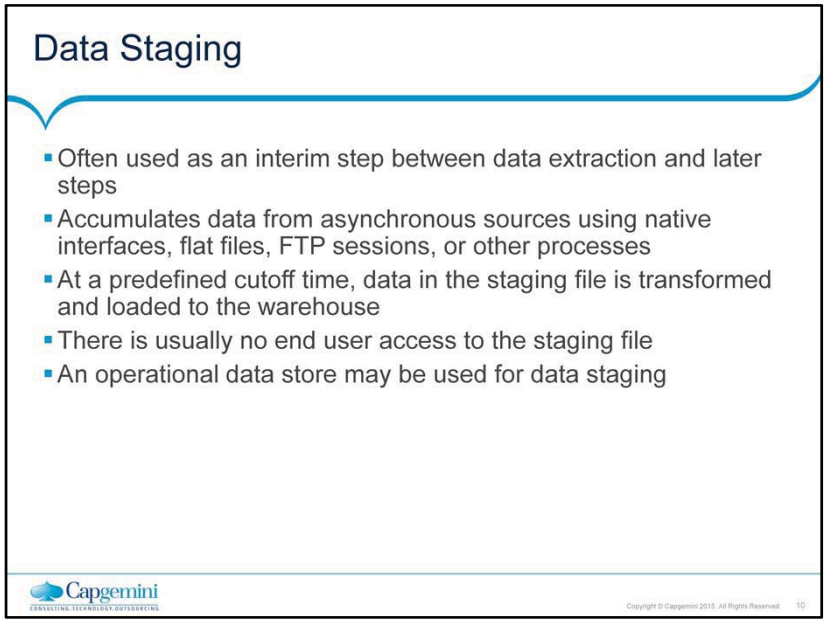
 8

Change Data Capture

- Based on the Publisher/Subscriber model.
- Publisher
  - Identifies the source tables from which the change data needs to be captured
  - Captures the change data and stores it in specially created change tables
  - Allows the subscribers controlled access to the change data
- Subscriber
  - Subscriber needs to know what change data it is interested in
  - It creates a subscriber view to access the change data to which it has been granted access by the publisher

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

 9



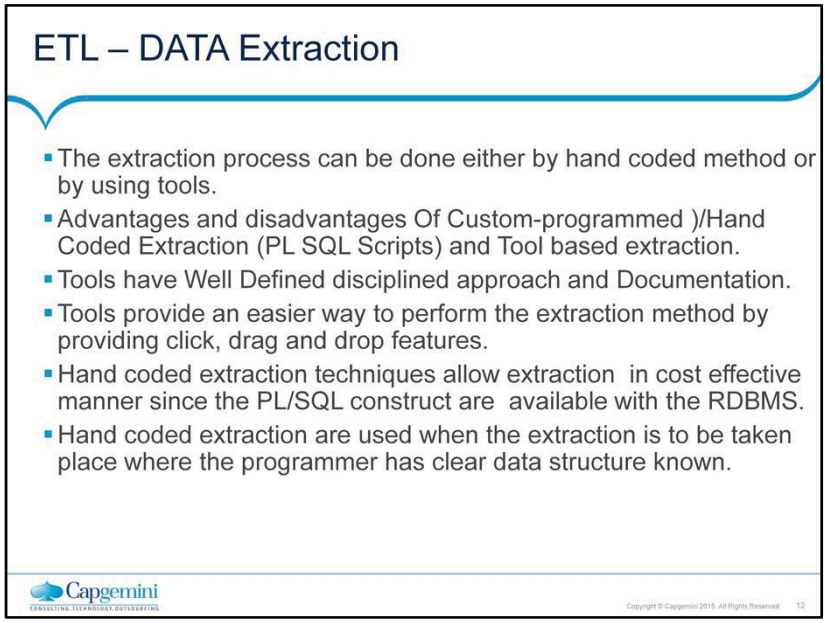
Data staging is used in cleansing, transforming, and integrating the data.

Reasons for “Dirty” Data

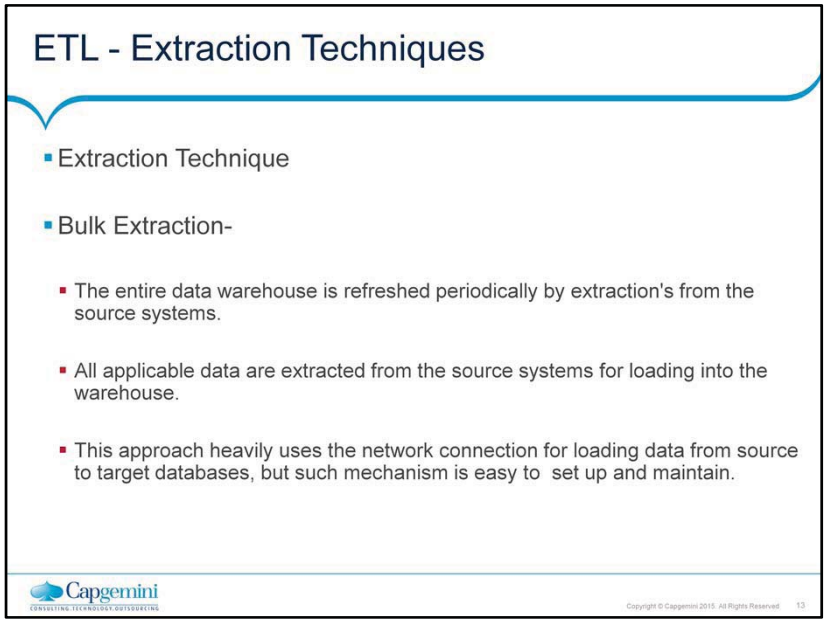
- Dummy Values
- Absence of Data
- Multipurpose Fields
- Cryptic Data
- Contradicting Data
- Inappropriate Use of Address Lines
- Violation of Business Rules
- Reused Primary Keys,
- Non-Unique Identifiers
- Data Integration Problems

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

 11



Though the extraction process can be done in either of the methods i.e either by hand coded methods or by using the tools. Tool based extraction have a well defined approach with a better documentation and it also makes the extraction process easier by a simple click, drag and drop features that are more user-friendly to the programmers.



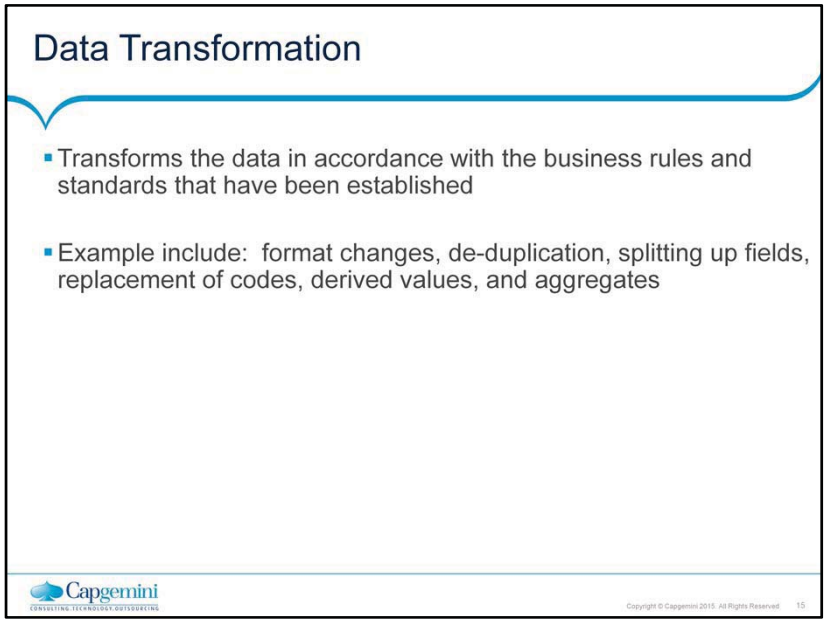
Bulk extraction needs the entire data warehouse to be refreshed periodically in which the entire data which is there in the data warehouse and the data to be loaded in to the warehouse are loaded once again in to the warehouse which uses heavy network traffic. But this mechanism is much easier to set up and maintain .

Data Extraction

- Capture of data from Source Systems
- Important to decide the frequency of Extraction
- Sometimes source data is copied to the target database using the replication capabilities of standard RDBMS (not recommended because of “dirty data” in the source systems)

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

 14



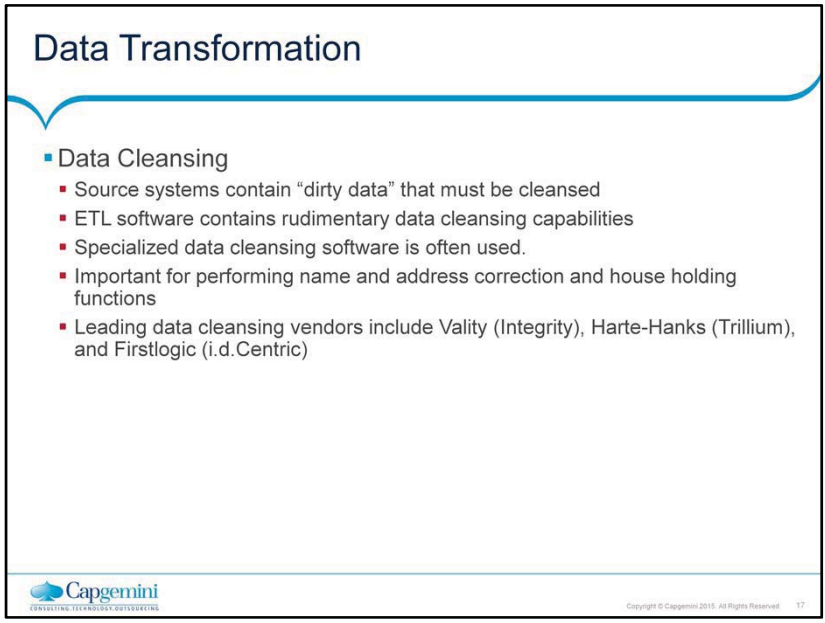
Aggregates, such as sales totals, are often precalculated and stored in the warehouse to speed queries that require summary totals.

Data Transformation

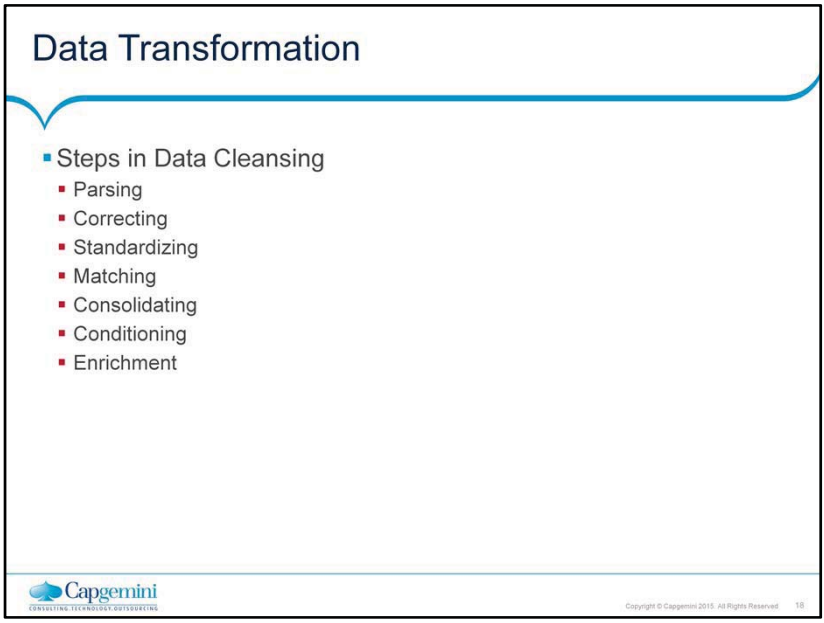
- Validating
  - Process of ensuring that the data captured is accurate and transformation process is correct
  - E.g. Date of Birth of a Customer should not be more than today's date

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

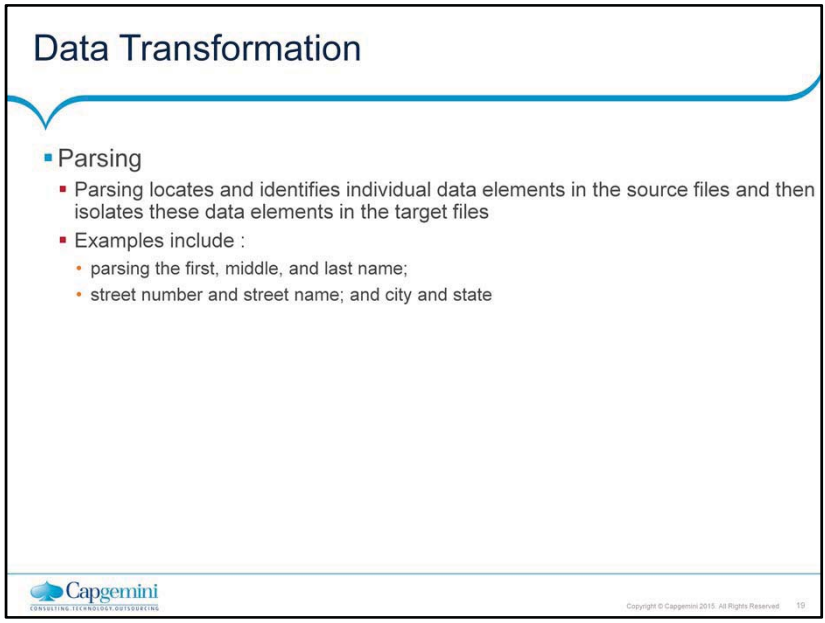
 16



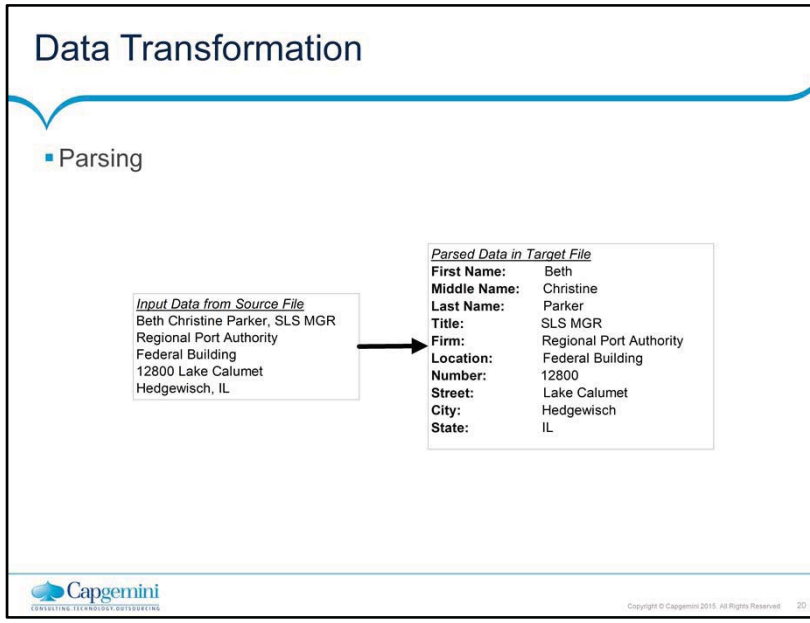
Data cleansing is critical to customer relationship management initiatives.

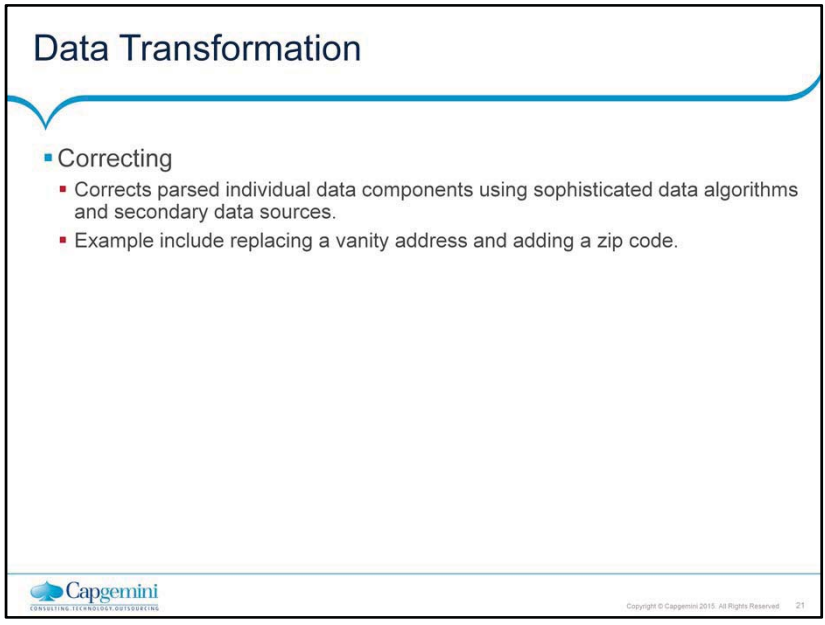


A good example to use is cleansing customer data. Most students can identify with receiving multiple copies of the same catalog because the company is not doing a good data cleansing job.

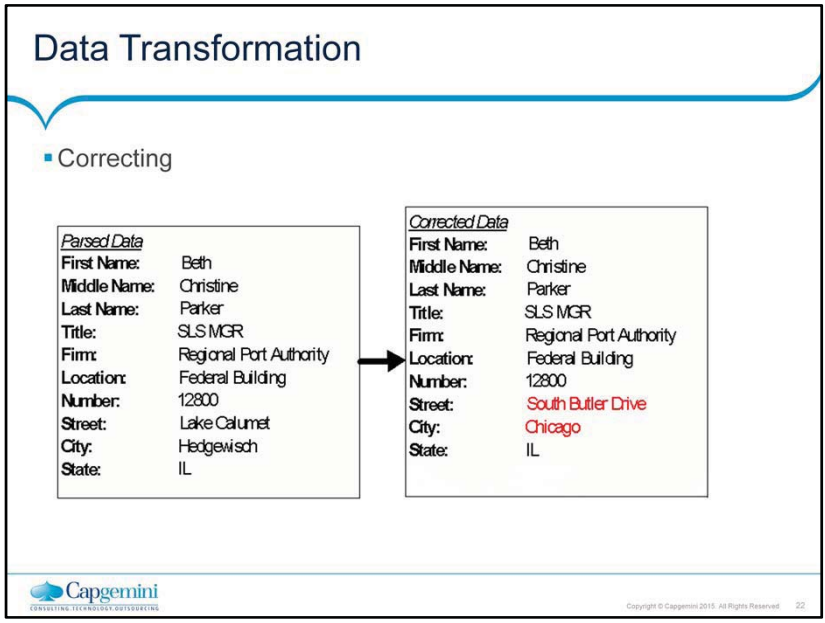


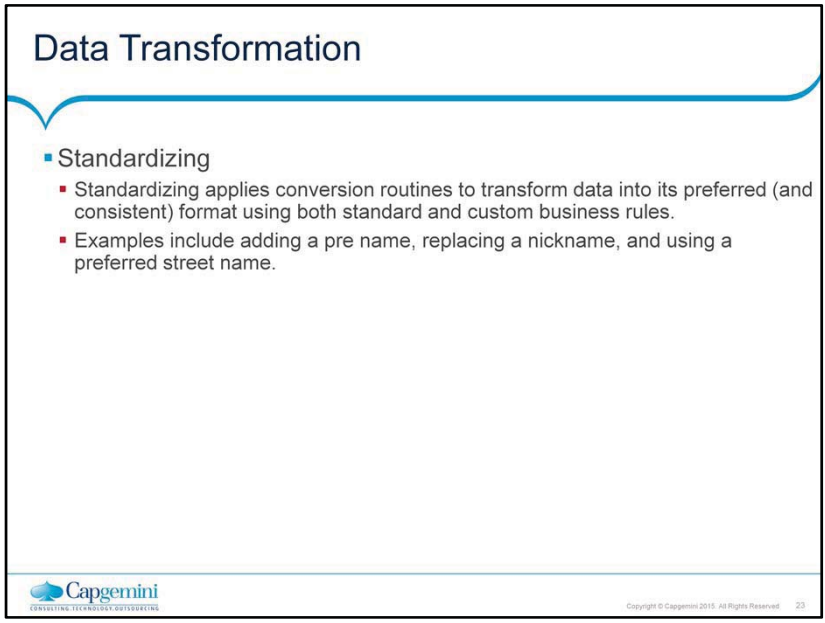
The record is broken down into atomic data elements.


Data Transformation
Parsing
Input Data from Source File
Beth Christine Parker, SLS MGR
Regional Port Authority
Federal Building
12800 Lake Calumet
Hedgewisch, IL
Parsed Data in Target File
First Name: Beth
Middle Name: Christine
Last Name: Parker
Title: SLS MGR
Firm: Regional Port Authority
Location: Federal Building
Number: 12800
Street: Lake Calumet
City: Hedgewisch
State: IL
Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING
Copyright © Capgemini 2015. All Rights Reserved 20

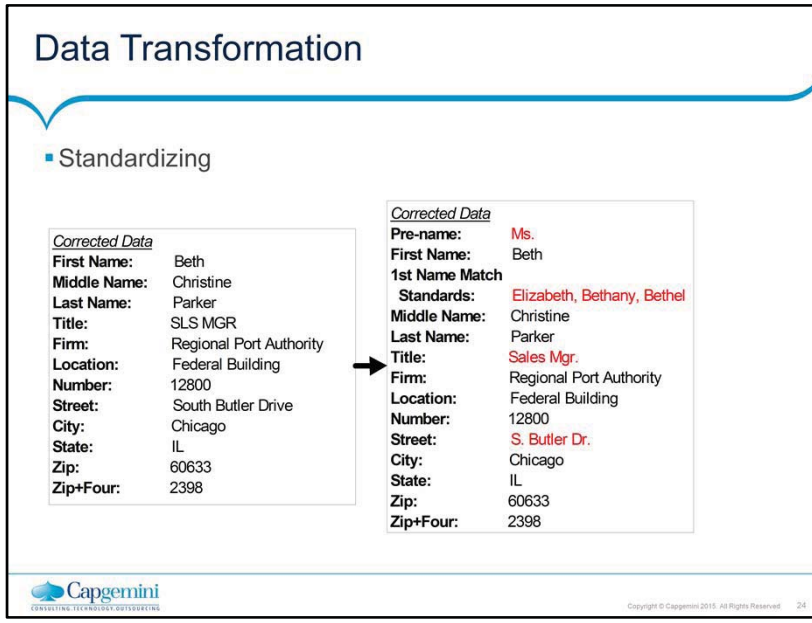


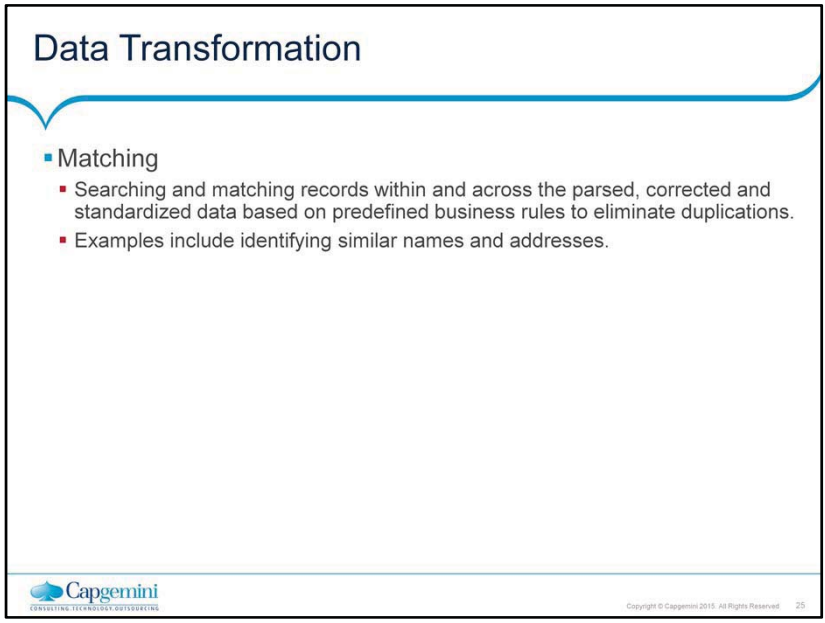
External data, such as census data, is often used in this process.


Data Transformation
Correcting
Parsed Data
First Name: Beth
Middle Name: Christine
Last Name: Parker
Title: SLS MGR
Firm: Regional Port Authority
Location: Federal Building
Number: 12800
Street: Lake Calumet
City: Hedgewisch
State: IL
Corrected Data
First Name: Beth
Middle Name: Christine
Last Name: Parker
Title: SLS MGR
Firm: Regional Port Authority
Location: Federal Building
Number: 12800
Street: South Butler Drive
City: Chicago
State: IL
Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING
Copyright © Capgemini 2015. All Rights Reserved 22

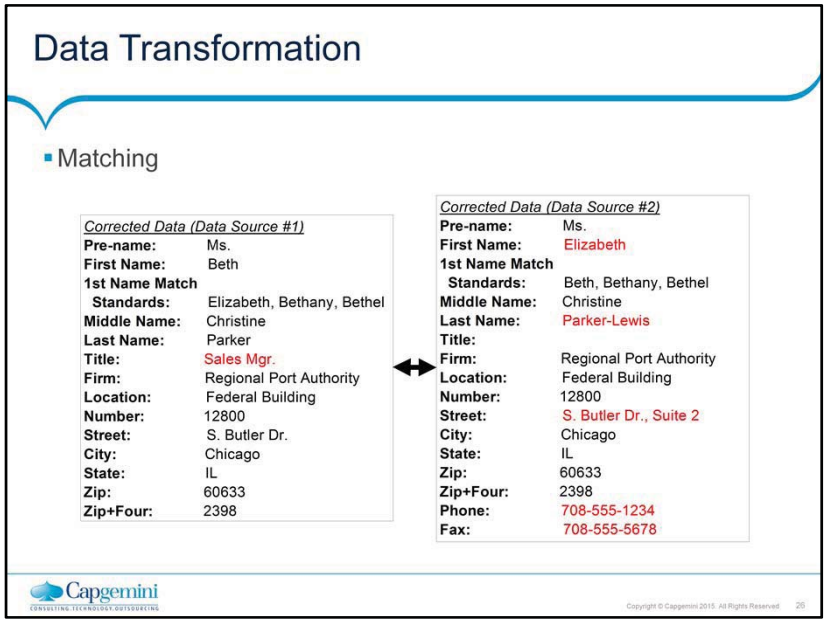


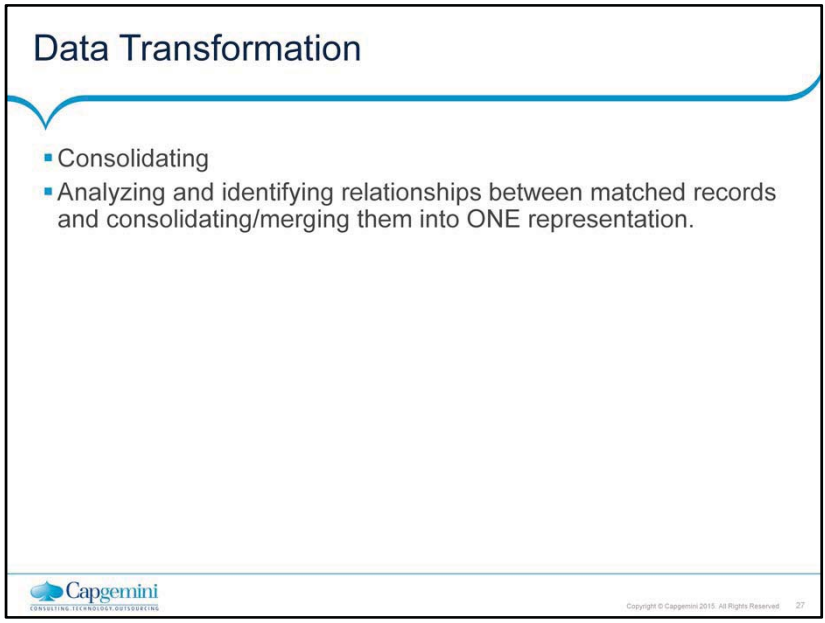
Companies decide on the standards that they want to use.


Data Transformation
Standardizing
Corrected Data
First Name: Beth
Middle Name: Christine
Last Name: Parker
Title: SLS MGR
Firm: Regional Port Authority
Location: Federal Building
Number: 12800
Street: South Butler Drive
City: Chicago
State: IL
Zip: 60633
Zip+Four: 2398
Corrected Data
Pre-name: Ms.
First Name: Beth
1st Name Match Standards: Elizabeth, Bethany, Bethel
Middle Name: Christine
Last Name: Parker
Title: Sales Mgr.
Firm: Regional Port Authority
Location: Federal Building
Number: 12800
Street: S. Butler Dr.
City: Chicago
State: IL
Zip: 60633
Zip+Four: 2398
Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING
Copyright © Capgemini 2015. All Rights Reserved 24

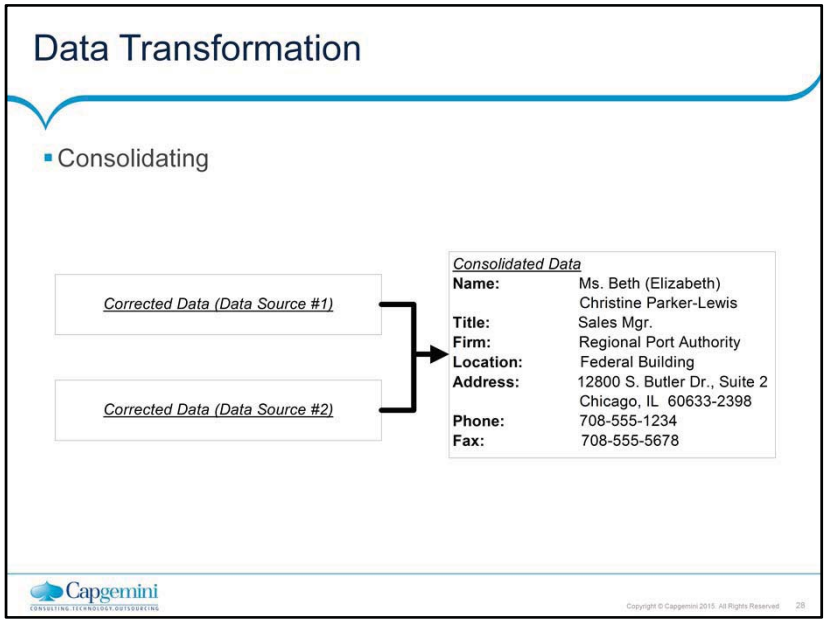


Commercial data cleansing software often uses AI techniques to match records.


Data Transformation
Matching
Corrected Data (Data Source #1)
Pre-name: Ms.
First Name: Beth
1st Name Match Standards: Elizabeth, Bethany, Bethel
Middle Name: Christine
Last Name: Parker
Title: Sales Mgr.
Firm: Regional Port Authority
Location: Federal Building
Number: 12800
Street: S. Butler Dr.
City: Chicago
State: IL
Zip: 60633
Zip+Four: 2398
Corrected Data (Data Source #2)
Pre-name: Ms.
First Name: Elizabeth
1st Name Match Standards: Beth, Bethany, Bethel
Middle Name: Christine
Last Name: Parker-Lewis
Title:
Firm: Regional Port Authority
Location: Federal Building
Number: 12800
Street: S. Butler Dr., Suite 2
City: Chicago
State: IL
Zip: 60633
Zip+Four: 2398
Phone: 708-555-1234
Fax: 708-555-5678
Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING
Copyright © Capgemini 2015. All Rights Reserved 26



All of the data are now combined in a standard format.

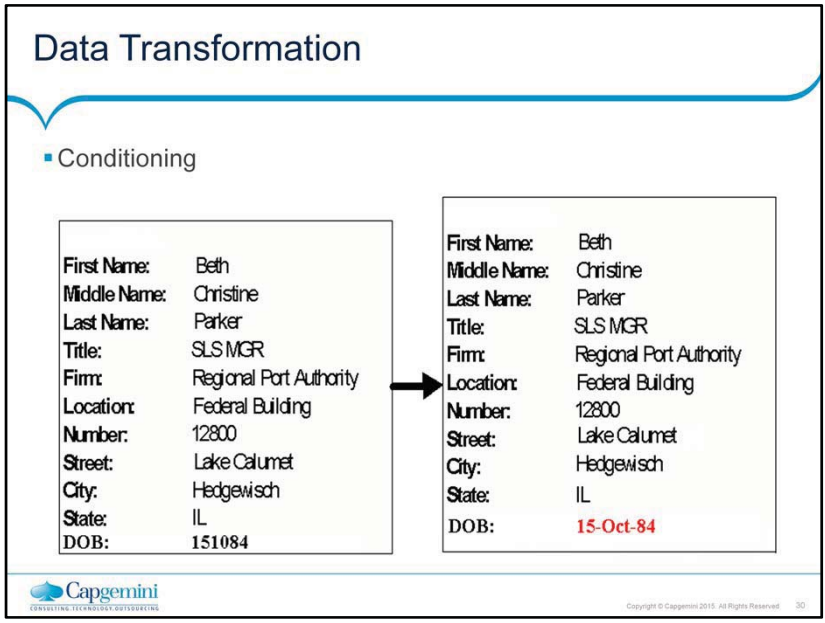

Data Transformation
Consolidating
Corrected Data (Data Source #1)
Corrected Data (Data Source #2)
Consolidated Data
Name: Ms. Beth (Elizabeth) Christine Parker-Lewis
Title: Sales Mgr.
Firm: Regional Port Authority
Location: Federal Building
Address: 12800 S. Butler Dr., Suite 2 Chicago, IL 60633-2398
Phone: 708-555-1234
Fax: 708-555-5678
Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING
Copyright © Capgemini 2015. All Rights Reserved 28

Data Transformation

- Conditioning
  - The conversion of data types from the source to the target data store (warehouse) -- always a relational database
  - Eg. OLTP Date stored as text (DDMMYY); DW format is Oracle Date type

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

 29

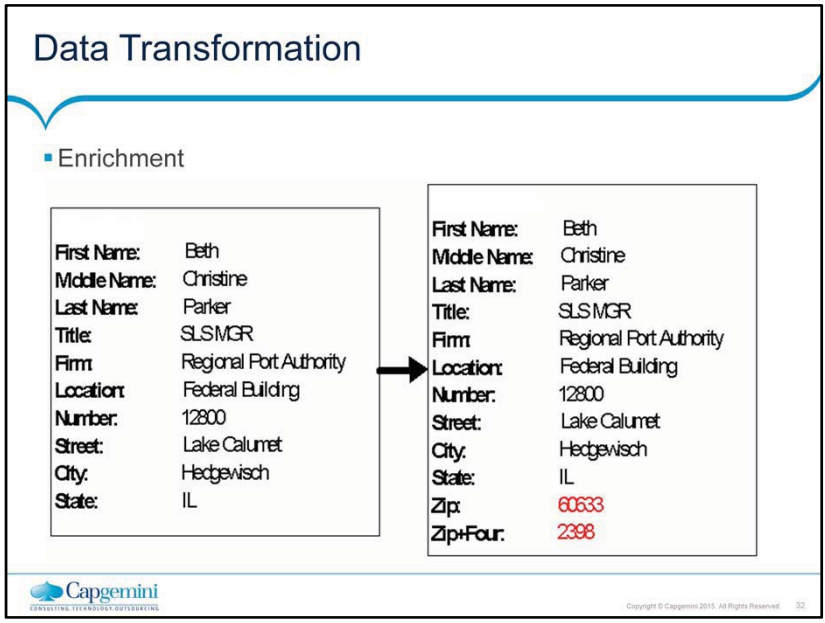

Data Transformation
Conditioning
First Name: Beth
Middle Name: Christine
Last Name: Parker
Title: SLS MGR
Firm: Regional Port Authority
Location: Federal Building
Number: 12800
Street: Lake Calumet
City: Hedgewisch
State: IL
DOB: 151084
First Name: Beth
Middle Name: Christine
Last Name: Parker
Title: SLS MGR
Firm: Regional Port Authority
Location: Federal Building
Number: 12800
Street: Lake Calumet
City: Hedgewisch
State: IL
DOB: 15-Oct-84
Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING
Copyright © Capgemini 2015. All Rights Reserved 30

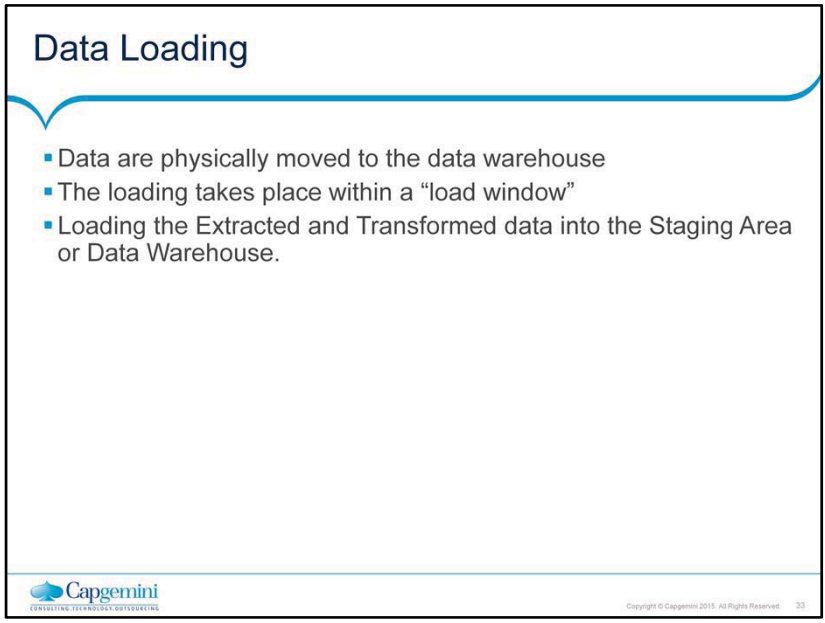
Data Transformation

- Enrichment
  - Adding/combining external data values, rules to enrich the information already existing in the data
  - E.g. If we can get a list that provides a relationship between Zip Code, City and State, then if a address field has Zip code 06905 it be safely assumed and address can be enriched by doing a lookup on this table to get Zip Code 06905 –> City Stamford –> State CT

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

 31


Data Transformation
Enrichment
First Name: Beth
Middle Name: Christine
Last Name: Parker
Title: SLS MGR
Firm: Regional Port Authority
Location: Federal Building
Number: 12800
Street: Lake Calumet
City: Hedgewisch
State: IL
First Name: Beth
Middle Name: Christine
Last Name: Parker
Title: SLS MGR
Firm: Regional Port Authority
Location: Federal Building
Number: 12800
Street: Lake Calumet
City: Hedgewisch
State: IL
Zip: 60633
Zip+Four: 2398
Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING
Copyright © Capgemini 2015. All Rights Reserved 32



Most loads involve only change data rather than a bulk reloading of all of the data in the warehouse.

Data Loading

- First time bulk load to get the historical data into the Data Warehouse
- Periodic Incremental loads to bring in modified data
- Design load strategy to using appropriate Slowly Changing Dimension type .
- The Loading window should be as small as possible
- Should be clubbed with strong Error Management process to capture the failures or rejections in the Loading process

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

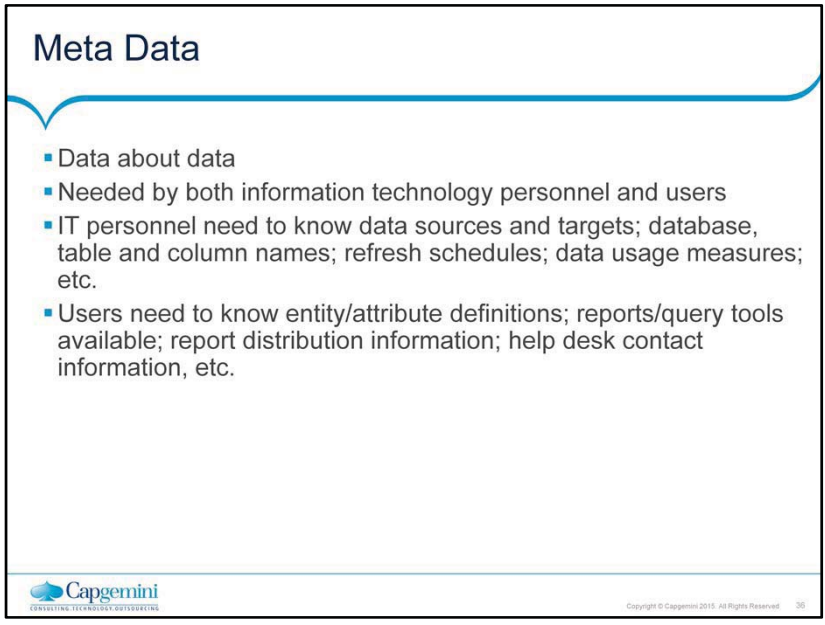
 34

Slowly Changing Dimension Types

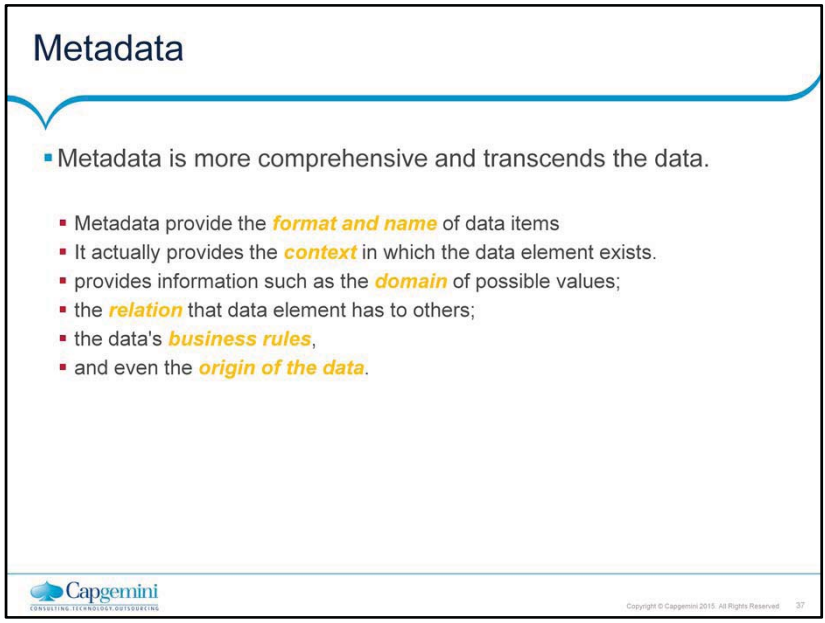
- Three types of slowly changing dimensions
  - Type 1
    - Updates existing record with modifications
    - Does not maintain history
  - Type 2
    - Adds new record
    - Maintain history
    - Maintains old record
  - Type 3:
    - Keep old and new values in the existing row
    - Requires a design change

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

 35

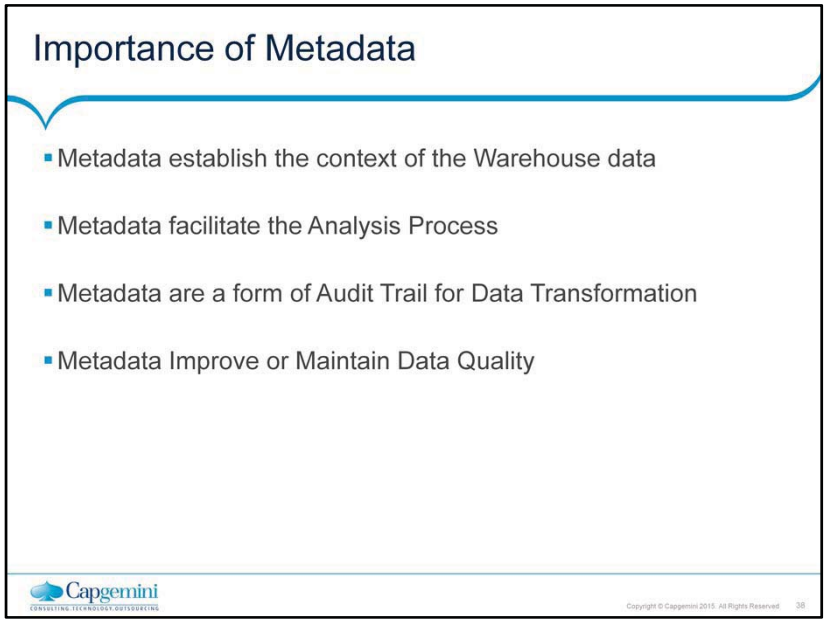


The importance of meta data is now realized, even though creating it is not glamorous work.



Metadata is the high level core internal document of the source code which runs as the lifeblood for a data warehouse.

Metadata not only describe the format and name but it provides details about the context I,e what is the need of the data item and what are the values that the data item can have, the relationship between the data elements ie whether the data element is found on other locations and how they are inter-linked to each other. Apart from the technical details It also holds the business rule. The origin of the data is so critical that the end user might like to trace back to the origin of the data which end user sees through the OLAP tools.



Importance of Metadata

Metadata establish the context of the Warehouse data
Metadata helps data warehouse administrators and users locate and understand data items, both in the source systems and in the warehouse data structures.

E.g.: The date 02/05/2010 could mean either May 2, 2010 or February 5, 2010 depending on the date convention used. Metadata describing the format of this date field could help determine the definite and unambiguous meaning of the data item.

Metadata facilitate the Analysis Process
Metadata must provide data warehouse end-users with the information they need to easily perform the analysis steps. It should thus allow users to quickly locate data that are in the warehouse.

Metadata should allow analysts to interpret data correctly by providing information about data formats and data definitions.

Metadata are a form of Audit Trail for Data Transformation
Metadata document the transformation of source data into warehouse data. Hence warehouse metadata must be capable of explaining how a particular piece of warehouse data was derived from the operational systems.

All business rules governing the transformation of data to new values or new formats are also documented as metadata.

Metadata Improve or Maintain Data Quality
Metadata can improve or maintain warehouse data quality through the definition of valid values for individual warehouse data items. Using a data quality tool prior to actual loading into the warehouse, the warehouse load images can be reviewed to check for compliance with valid values for key data items. Data errors are quickly highlighted for correction.

Metadata can be used as the basis for any error-correction processing that should be done if a data error is found. Error-correction rules are documented in the metadata repository and executed by program code on an as needed basis.

All business rules governing the transformation of data to new values or new formats are also documented as metadata.

Metadata Improve or Maintain Data Quality
Metadata can improve or maintain warehouse data quality through the definition of valid values for individual warehouse data items. Using a data quality tool prior to actual loading into the warehouse, the warehouse load images can be reviewed to check for compliance with valid values for key data items. Data errors are quickly highlighted for correction.

Metadata can be used as the basis for any error-correction processing that should be done if a data error is found. Error-correction rules are documented in the metadata repository and executed by program code on a need basis.

Feature of ETL Tools

- Support data extraction, cleansing, aggregation, reorganization, transformation, and load operations
- Generate and maintain centralized metadata
- Filter data, convert codes, calculate derived values, map source data fields to target data fields
- Automatic generation of ETL programs
- Closely integrated with RDBMS
- High speed loading of target data warehouses using Engine-driven ETL Tools

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

 40

Advantages of using ETL Tools

- GUI based design of jobs – ease of development and maintenance
- Generation of directly executable code
- Engine driven technology is fast, efficient and multithreaded
- In-memory data streaming for high-speed data processing
- Products are easy to learn and require less training

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

 41

Advantages of using ETL Tools

- Automatic generation and maintenance of open, extensible metadata
- Support for multiple data formats and platforms
- Large number of vendor supplied data transformation objects

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

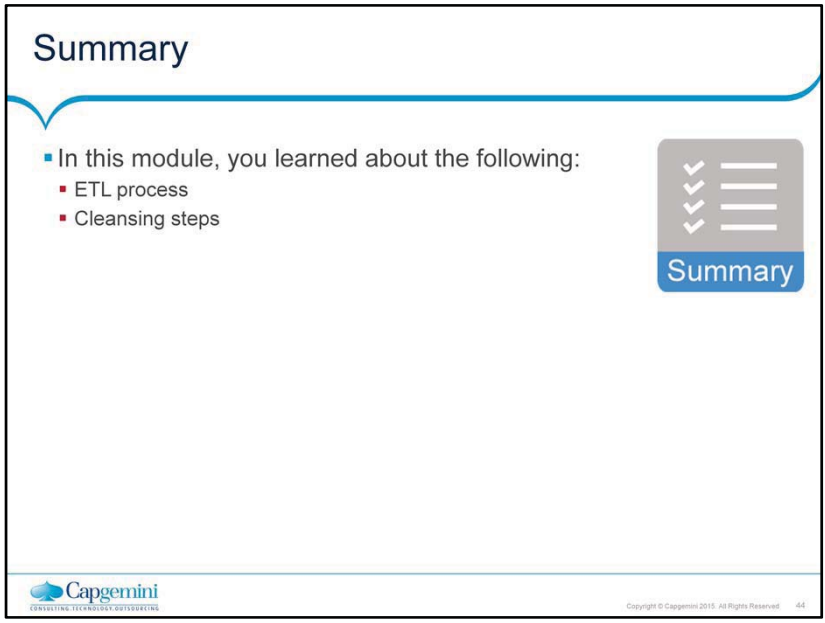
 42

Example of ETL requirements

- Integration of masters across different systems
  - E.g. State code AP could mean Andhra Pradesh in one system while it could mean Arunachal Pradesh in another
- De-duplication of data from different systems
  - E.g. State Karnataka could be represented as KA in one system and KN in another system
- Mapping of old codes to Data Warehouse codes
- Data Cleansing - Changing to upper case, assigning defaults to unavailable data elements

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

 43



Add the notes here.

ETL Basics

Lesson 3: Operational Considerations

Lesson Objectives

- On completion of this lesson on ETL basics, you will be able to understand:
  - Handling Exceptions in ETL
  - Notifications and Alerts in ETL
  - Recovery and Restartability

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

 2

ETL Testing Considerations

- UNIT Testing (ensures that each component within the system successfully performs its individual responsibility when executed individually.)
  - Checking extraction rules
  - Transformation validation
  - Target system data integrity
  - Checking input data validation
  - Test the error-handling logic
  - Test slowly changing dimension implementation by checking the integrity of surrogate keys
  - Test Notifications/Warnings/Error messages
- Integration Testing (ensure seamless run of the entire process within an application, or a specific stage with an eye on the details of each of the steps/modules, capturing the responses as the data moves across the system.)

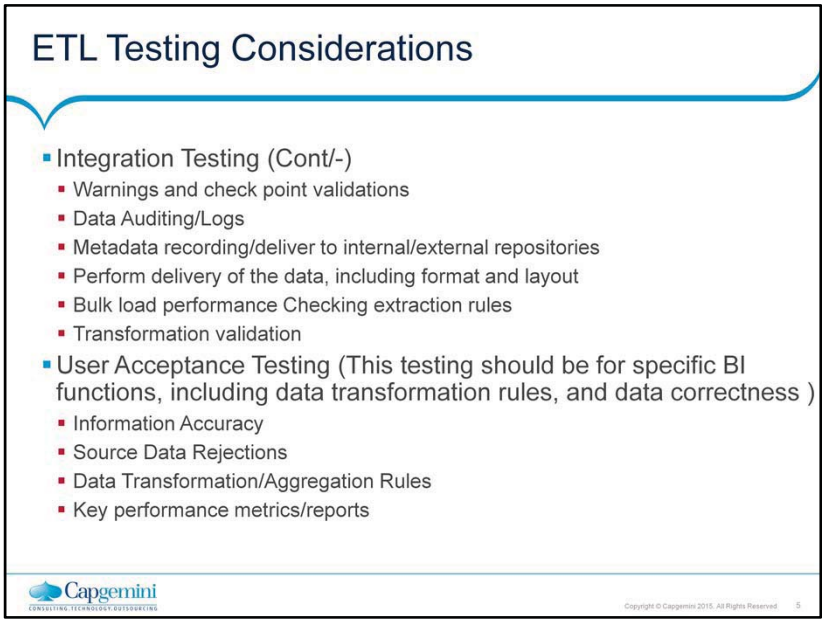
Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

 3

ETL Testing Considerations (contd..)

- Successful extraction of data
- Order of Extraction
- Application and validation of transformation logic
- Order of precedence in which various algorithms are applied (phasing of ETL streams)
- Rejects based on applied algorithms
- Recovery and Restart
- Proper generation of the code
- Proper generation of surrogated keys in conjunction with processing the order of precedence
- Error handling
- Scheduling
- Job triggers
- Job dependencies
- Alerts and notification

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING


ETL Testing Considerations
Integration Testing (Cont/-)
Warnings and check point validations
Data Auditing/Logs
Metadata recording/deliver to internal/external repositories
Perform delivery of the data, including format and layout
Bulk load performance Checking extraction rules
Transformation validation
User Acceptance Testing (This testing should be for specific BI functions, including data transformation rules, and data correctness)
Information Accuracy
Source Data Rejections
Data Transformation/Aggregation Rules
Key performance metrics/reports
Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING
Copyright © Capgemini 2015. All Rights Reserved. 5

Exception Handling

- Exception Handling deals with any abnormal termination, unacceptable event or incorrect data that can impact the data flow or accuracy of data in the warehouse/mart.
- Exceptions in ETL could be classified as Data Related Exceptions and Infrastructure Related Exceptions.
- The process of recovering or gracefully exiting when an exception occurs is called exception handling.
- Data related exceptions are caused because of incorrect data format, incorrect value, incomplete data from the source system. This leads to Data validation exceptions and Data Rejects.
- The process of handling the Data Rejects is called Data Reprocessing.

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

 6

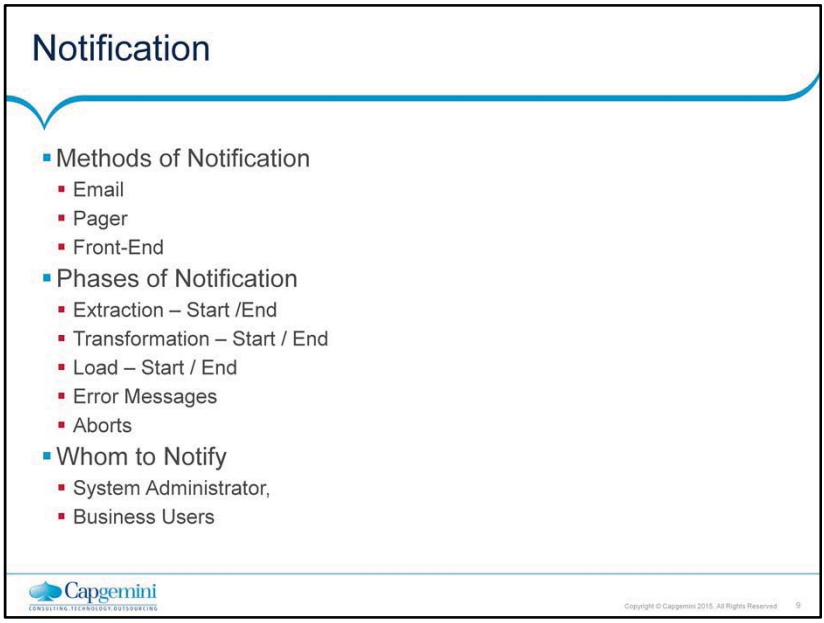
Exception Handling

- Infrastructure related exceptions are caused because of issues in the Network , the Database and the Operating System.
- Common Infrastructure exceptions are FTP failure, Database connectivity failure, File system full etc.
- The data related exceptions are usually documented in the requirements, if not they must be because if the data related exceptions are not handled they lead to inaccurate data in the warehouse/mart.
- We also keep a threshold of maximum number of validation or reject failures allowed per load.
- Any value above the threshold would mean the data would be too inaccurate due to too many rejections.

Exception Handling

- There is one more exception which is the presence of inaccurate or incorrect data in the warehouse. This could happen due to
  - Incorrect requirement or missed, leading to incorrect ETL.
  - Incorrect interpretation of requirements leading to incorrect ETL.
  - Uncaught coding defects.
  - Incorrect data from source.
- The process of Correction of the data already loaded in the warehouse involves fixing the data already loaded and also preventing the inaccuracy to persist in the future.

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING


Notification
Methods of Notification
Email
Pager
Front-End
Phases of Notification
Extraction – Start /End
Transformation – Start / End
Load – Start / End
Error Messages
Aborts
Whom to Notify
System Administrator,
Business Users
Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING
Copyright © Capgemini 2015. All Rights Reserved. 9

Recovery & Restartability

- Design of Rollback and Recovery Procedures
  - Rollback and Recovery Procedures define the strategy for handling load failures.
  - This includes recommendations on whether milestone points and staging are required for restarts.
  - The concept of rollback and recovery of ETL processes needs to be considered during the preliminary design stage as it affects all ETL jobs.
  - For long ETL processing blocks such as a data warehouse load the end to end process may take hours or even days to run.
  - It is important to have built in restartability to recover from fatal errors during the processing cycle.

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

 10

Recovery & Restartability

- Milestone Recovery
  - The simplest form of recovery is to introduce vertical bands into the ETL processing cycle.
  - A vertical band is a set of jobs that run together and when complete they arrive at a milestone point.
  - If the following block of jobs fail it can be restarted from this milestone.
  - A milestone point requires some type of data staging, with the data being placed in temporary files or tables where they can be extracted by the next block.

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

 11

Recovery & Restartability

- The best practice for complex data loads is to have 3 vertical bands: Data Sourcing, Data Mediation & Quality/Transformation and Data Load.
  - Data Sourcing involves retrieving the data from sources and delivering it to files or tables with some type of time stamping to allow for time based processing.
  - This data undergoes as little transformation as possible and once delivered it is stable.
  - This milestone point means the Data Transformation block can be started and restarted without having the source data affected by user transactions.

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

 12

Recovery & Restartability

- Data Mediation & Quality/Transformation covers a large area of data conversion, cleansing, enrichment, aggregation, etc.
- This step benefits from having ETL patterns that describe common transformations as shown in the sections below.

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

 13

Recovery & Restartability

- Data Load involves delivering the final data to the target database.
  - This band holds as little of the transformation logic as possible.
  - It is focused on achieving a robust database update by controlling transaction sizing and trapping database rejects.
  - Database updates are the most volatile part of the process due to the complexity of RDBMS communications and the difficulty most ETL engines have with correctly rolling back and restarting a failed update.

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

 14

Recovery & Restartability

- Individual Job Recovery
  - In many instances it is possible to recover from a fatal error by restarting the job that failed and continuing the ETL cycle from that point.
  - Many ETL and scheduling tools provide the functionality to automate this process.
  - It may be necessary for production support to first investigate the problem and fix it before the automated recovery begins.
  - This is a short cut to restarting from previous milestone points.

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

 15

Recovery & Restartability

- Individual Job Recovery
  - It is usually easy to do job recovery on the Sourcing and Transformation bands as these typically stage the data to temporary files or tables.
  - Restarting an individual job recreates the target output of these jobs without rollback problems.
  - In these cases the ETL scheduling tool can be used to restart the sequence from the correct point.

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

 16

Recovery & Restartability

- Individual Job Recovery
  - The Data Load band is the most difficult for rollback and recovery as a job may fail in the process of updating a database.
  - If the update is an insert, or an update to an aggregated table then it is difficult to determine how many rows of stage data have already been processed.
  - A simple job restart may result in duplicate rows or duplicate increases to aggregate results.
  - For a Data Load job it may be possible to restart the job or it may be necessary to build full table rollback into a job restart.
  - It is worth considering enhancing the design to assist with rollback – for example, a batch number could be added to transaction tables to facilitate deletion of partial or erroneous insertions.

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

 17

Restartability Matrix

Issue	Steps to Mitigate Impact on Restartability	Party Responsible for Ensuring Steps are Completed
Data in source table changes frequently	**Append source data with a datestamp, and store a snapshot of** source data in a backup schema until the session has completed successfully	Database Administrator (creates backup schema in repository) Data Integration Developer (ensures that session calls backup schema when session recovery is performed)
Mappings in certain sessions are dependent on data produced by mappings in other sessions	**Arrange sessions in a sequential batch; configure sessions to run only if previous sessions are completed successfully**	Data Integration Developer
Session uses the Bulk Loading parameter	If sessions fail frequently due to external problems (e.g., network downtime), reconfigure the session to normal load. Bulk loading bypasses the database log, making session unrecoverable	Data Integration Developer
Only the Informatica Administrator can recover or restart sessions	Configure the session to send an email to the Informatica administrator when a session fails	Data Integration Developer
Multiple sessions within a concurrent batch fail	Work with database administrator to determine when failed sessions should be recovered, and when targets should be truncated and entire session run again.	Data Integration Developer Database Administrator

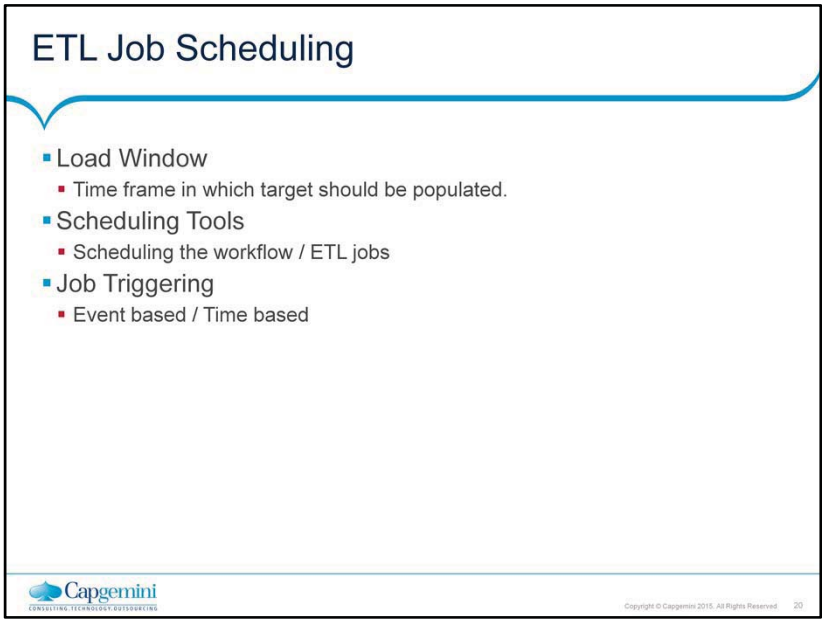
Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

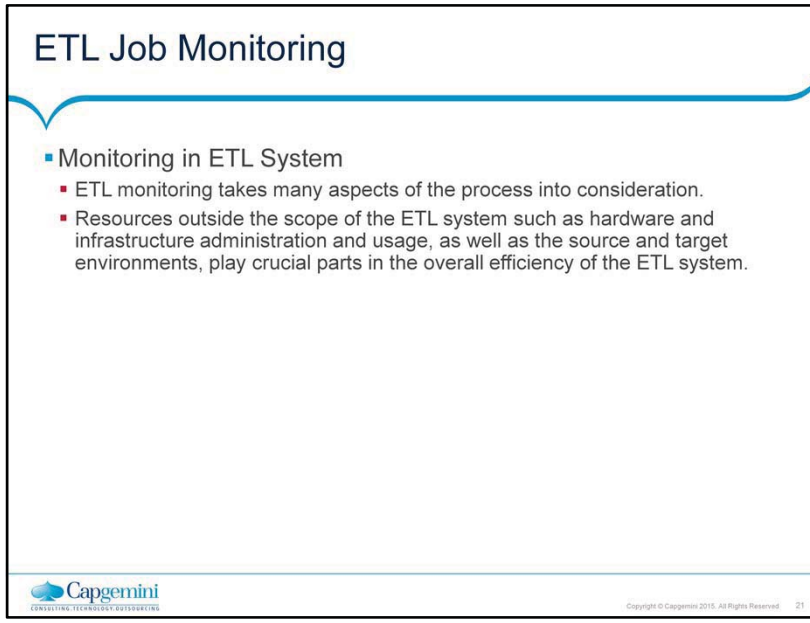
ETL Job Scheduling

- ETL Job Scheduling is an operational process which is required to determine the sequence and time of execution of the various data flows (Jobs/Mappings).
- ETL schedule is dependent on the following
  - Load Order
    - Order in which target data will be populated.
  - External Dependencies like
    - Timeframe of the source data availability.
    - Warehouse/Mart database Maintenance, like database backup time.
    - Operating System Maintenance, like file system backup time.
  - ETL's inter process flow dependencies like conformed dimensions ETL before the subject area specific dimension ETL, Dimension table ETL before the Fact Table ETL etc.

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

 19


ETL Job Scheduling
Load Window
Time frame in which target should be populated.
Scheduling Tools
Scheduling the workflow / ETL jobs
Job Triggering
Event based / Time based
Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING
Copyright © Capgemini 2015. All Rights Reserved. 20

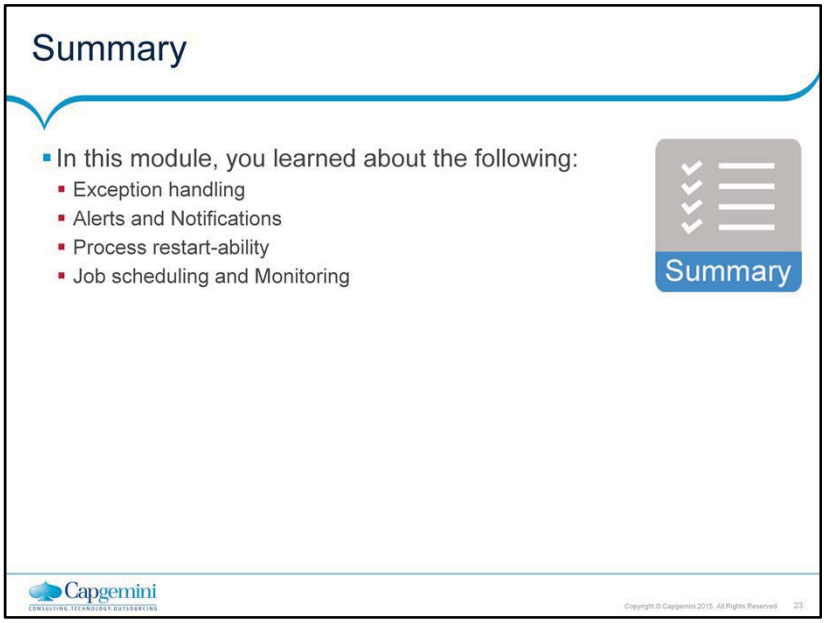

ETL Job Monitoring
Monitoring in ETL System
ETL monitoring takes many aspects of the process into consideration.
Resources outside the scope of the ETL system such as hardware and infrastructure administration and usage, as well as the source and target environments, play crucial parts in the overall efficiency of the ETL system.
Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING
Copyright © Capgemini 2015. All Rights Reserved. 21

ETL Job Monitoring

- Measuring ETL Specific Performance Indicators
  - Duration in seconds.
  - Rows processed per second.
  - Rows read per second.
  - Rows written per second.
  - Throughput.

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

 22



Add the notes here.

ETL Basics

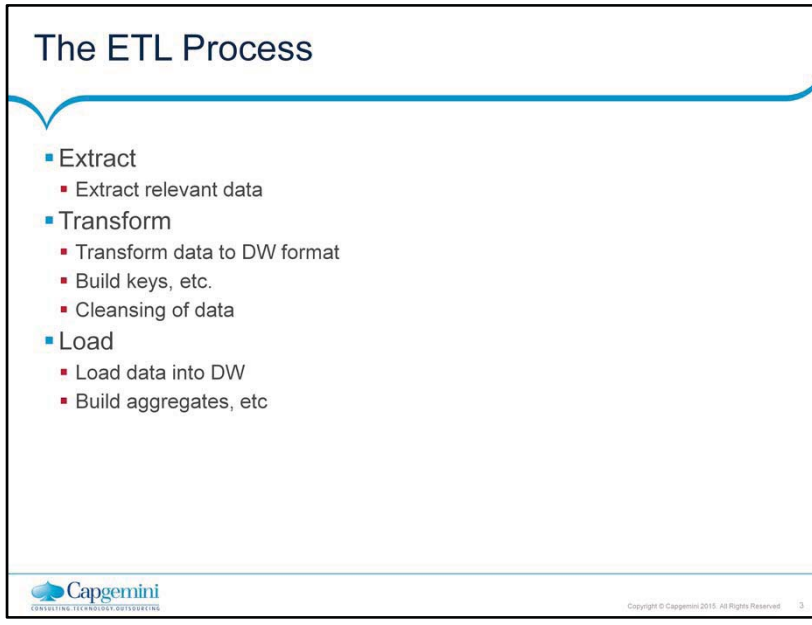
Lesson 4: ETL Tools

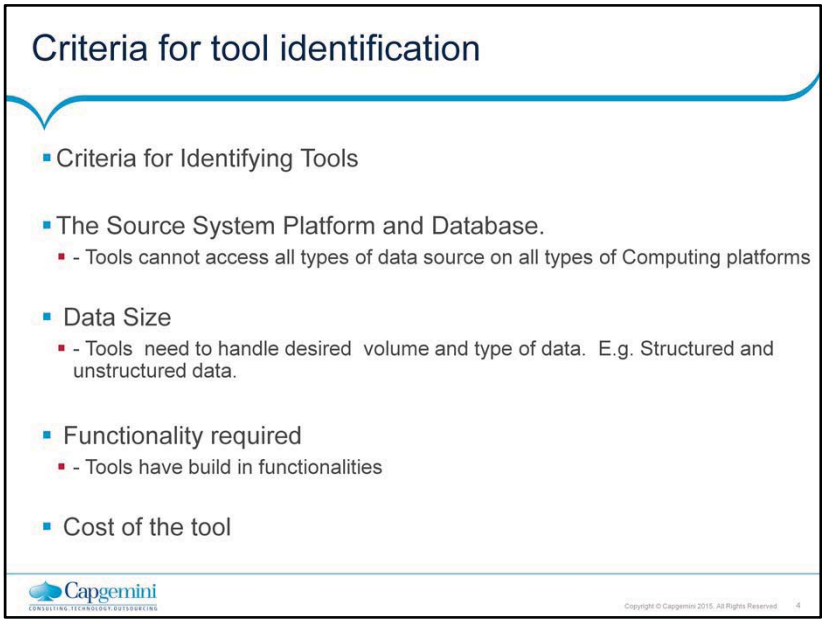
Lesson Objectives

- On completion of this lesson on ETL basics, you will be able to understand:
  - Consideration of ETL tool
  - Different ETL tools

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

 2


The ETL Process
Extract
Extract relevant data
Transform
Transform data to DW format
Build keys, etc.
Cleansing of data
Load
Load data into DW
Build aggregates, etc
Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING
Copyright © Capgemini 2015. All Rights Reserved 3



Various tools are available for extraction process. Selecting a suitable tools becomes an important criteria for the success of the warehouse implementation.

The tools should be capable of selecting the data from various sources without user interface. Connectors are available for extracting data from various sources from various computing platforms.

Some of the tools have built-in extraction which reduces the manual coding activity.

SAMPLE Extraction Tools

- Extraction Tools include
- Apertus Carleton. Passport
- Evolutionary Technologies. ETL Extract.
- Platinum. InfoPump

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

 5

These are the some of the Industry standard tools which are used for extracting data from either single or multiple sources.

Carleton, Passport
Users enter extraction and transformation parameters, and Data is filtered against domains and ranges of legal values.

Evolutionary Technology (ETI), EXTRACT
Users write transformation rules. Data is filtered against domains and ranges of legal values and compared to other data structures

Platinum, InfoPump
A data pump product designed to extract data from several mainframe and client server platforms, perform some filtering and transformation, and distribute and load to another mainframe platform database. Requires InfoHub for most services.

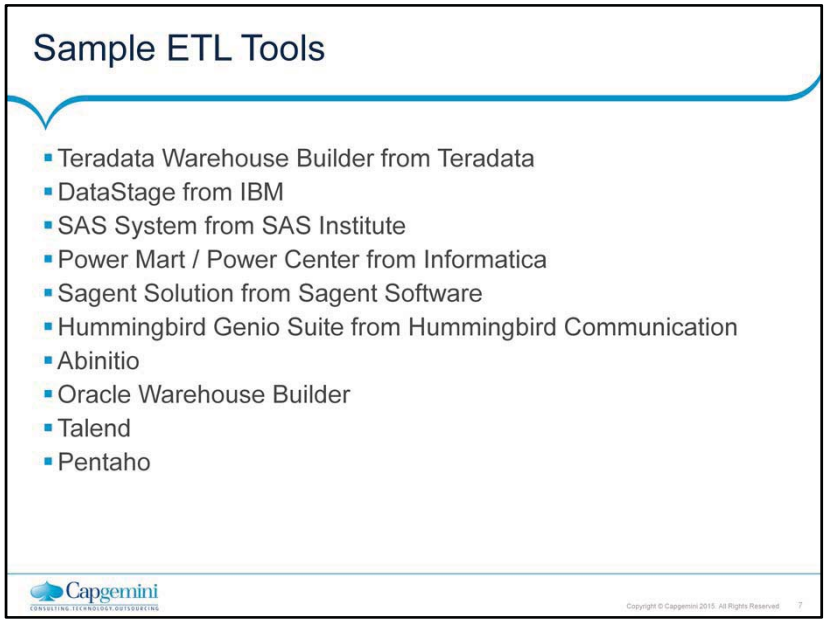
The extraction uses custom code modules. This is a client/server based tool.

In addition there some more tools available in the market such as
- Information Discovery ie., IDI
  Low-end rule discovery

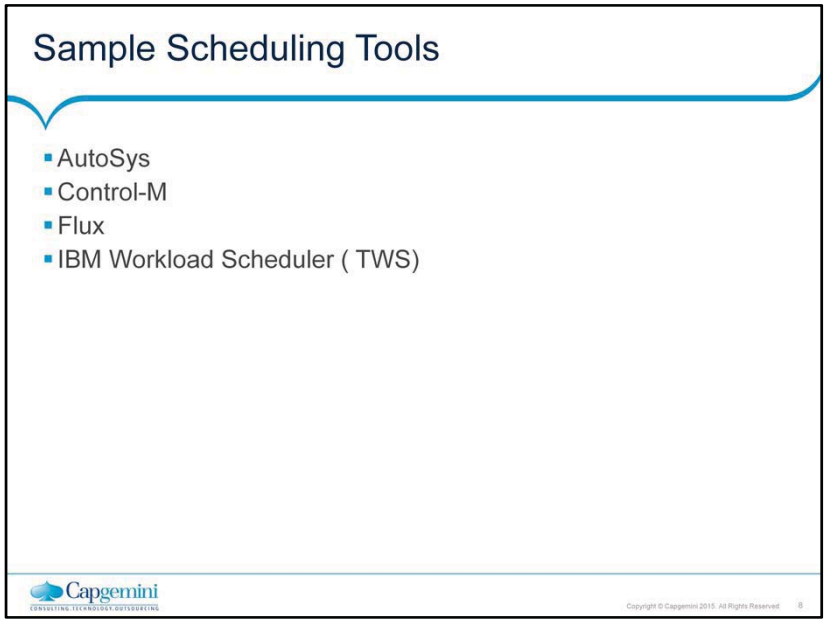
Oracle, Symmetric Replicator
A data replication product designed to extract data from several platforms, perform some filtering and transformation, and distributes and loads data in to one or more database or databases.

In addition there are some more tools available in the market such as
- Information Discovery, IDI
  Low-end rule discovery

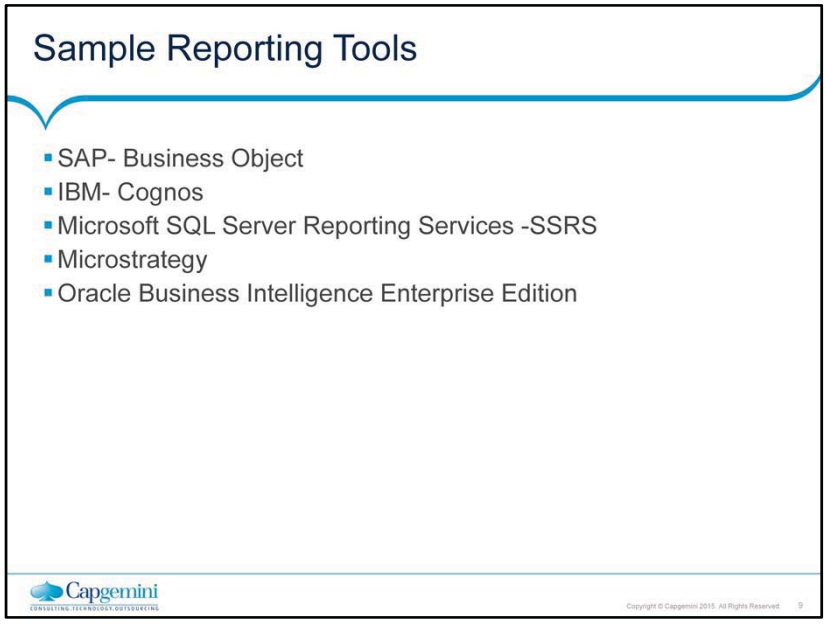
Oracle, Symmetric Replicator
A data replication product designed to extract data from several platforms, perform some filtering and transformation, and distributes and loads data in to one or more database or databases.



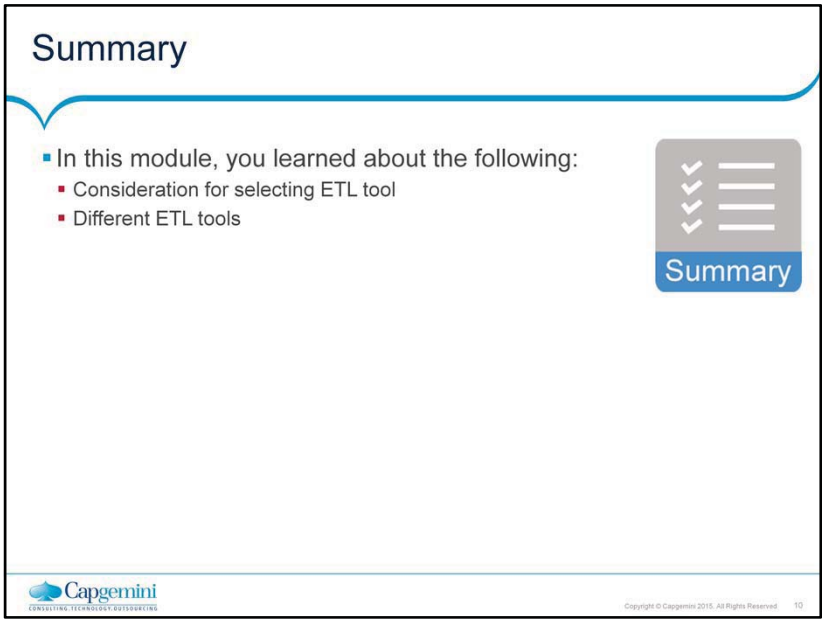
You might go to the vendors' web sites to find a good demo to show your students.



You might go to the vendors' web sites to find a good demo to show your students.



You might go to the vendors' web sites to find a good demo to show your students.



Add the notes here.