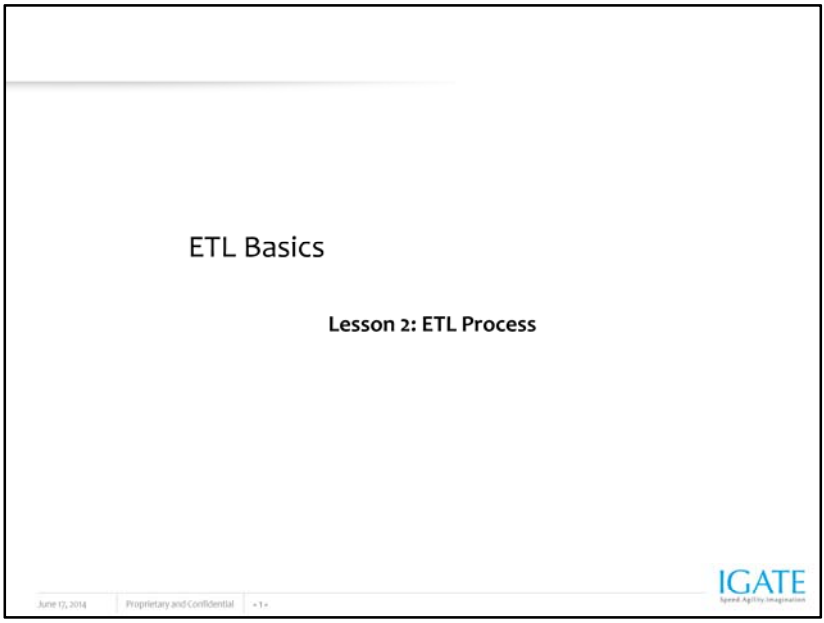
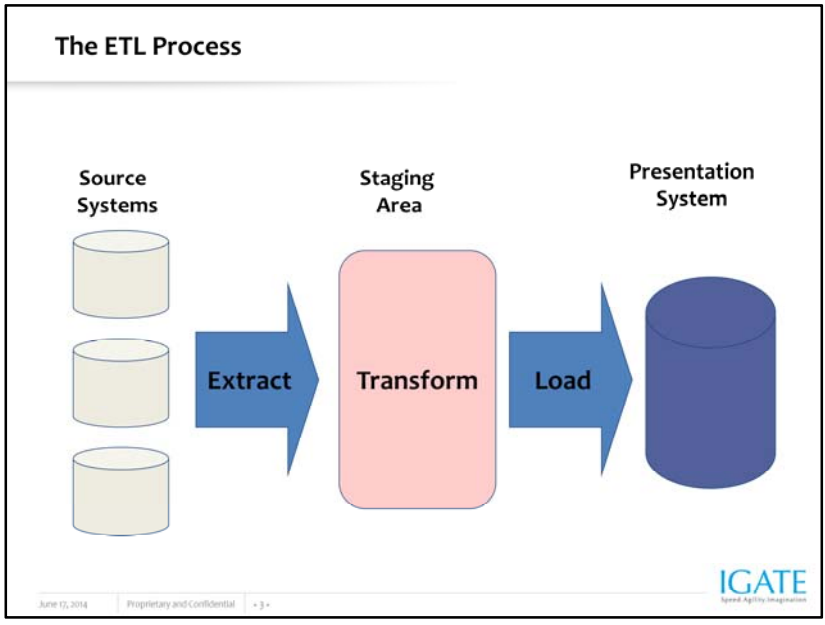

ETL Basics
Lesson 2: ETL Process
June 17, 2014
Proprietary and Confidential
IGATE

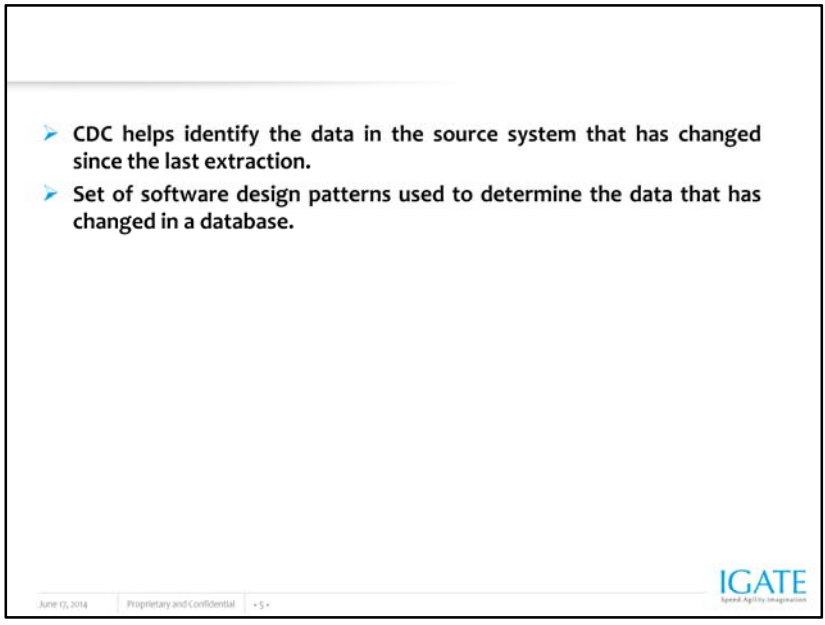
Lesson Objectives

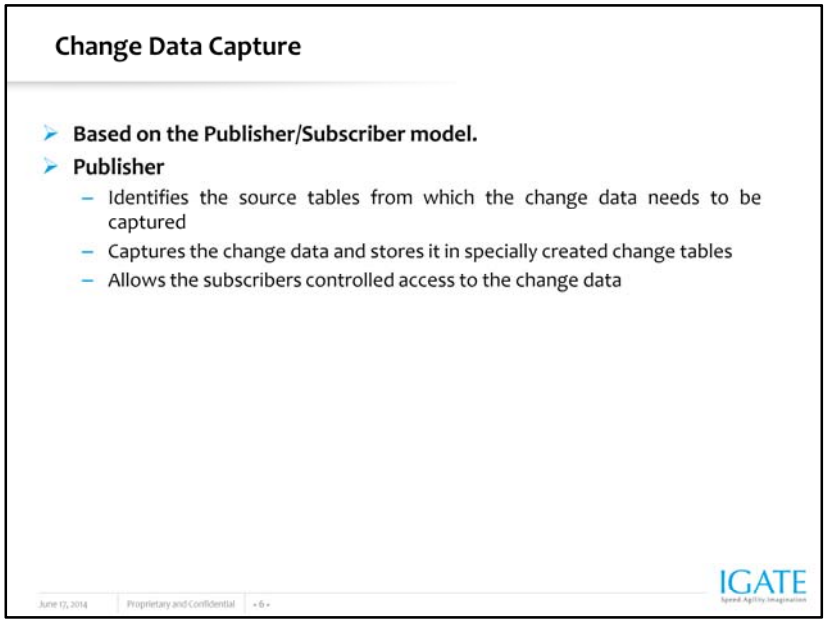
- **On completion of this lesson on Data Modeling, you will be able to understand:**
  - The ETL process
  - The steps in Data Cleansing

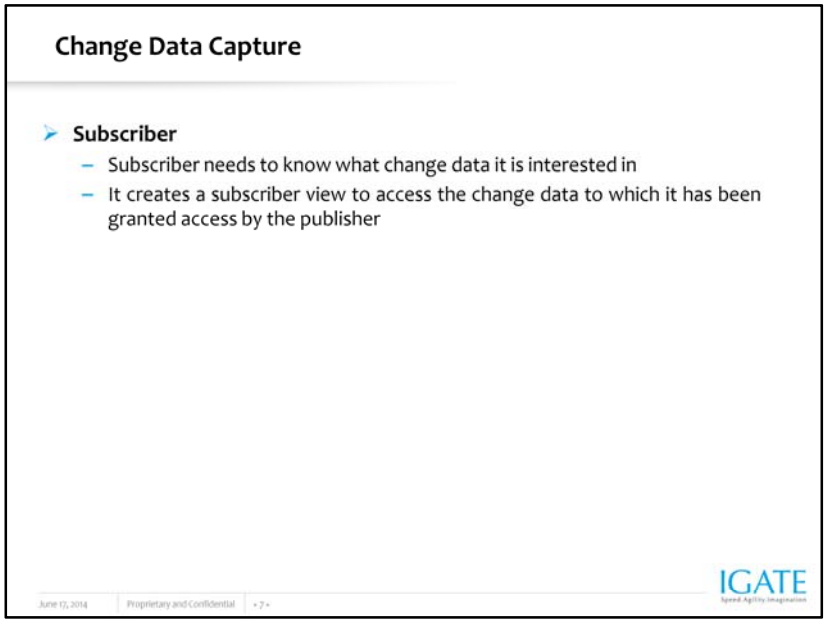

The ETL Process
Source Systems
Staging Area
Presentation System
Extract
Transform
Load
IGATE

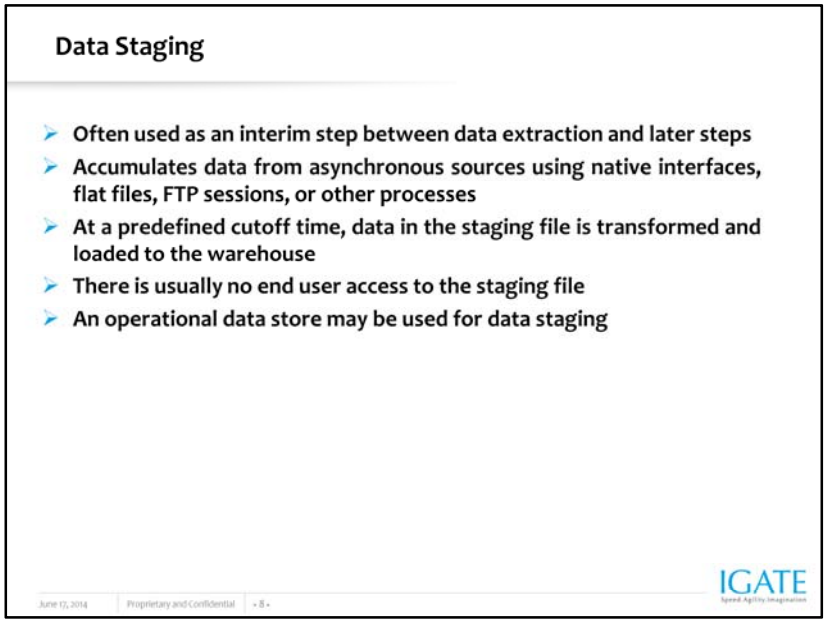
Change Data Capture

- 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).

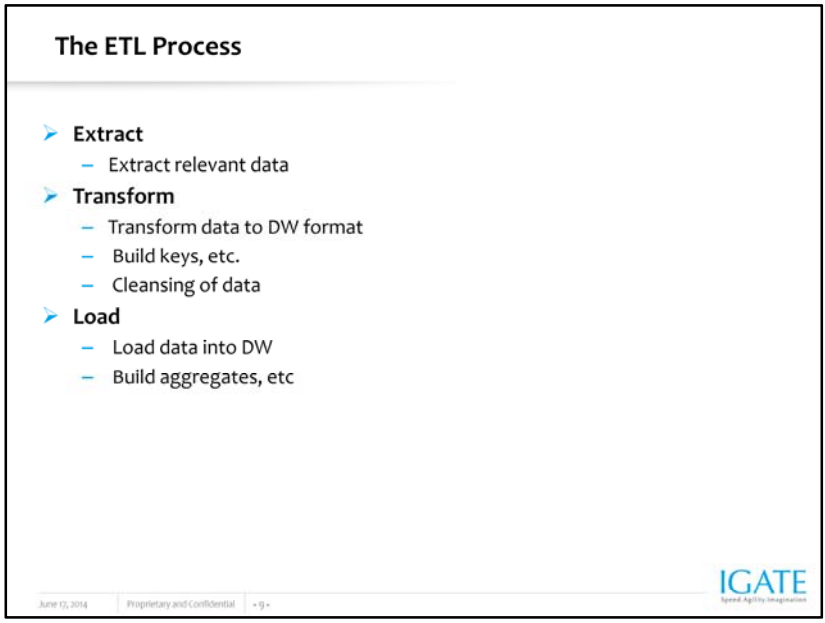

CDC helps identify the data in the source system that has changed since the last extraction.
Set of software design patterns used to determine the data that has changed in a database.
June 17, 2014
Proprietary and Confidential
IGATE

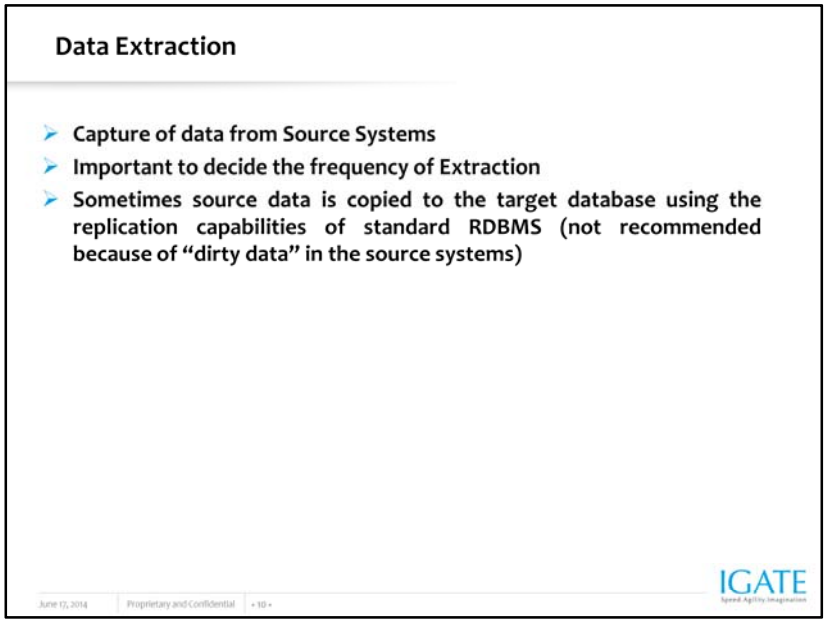

Change Data Capture
Based on the Publisher/Subscriber model.
Publisher
Identifies the source tables from which the change data needs to be captured
Captures the change data and stores it in specially created change tables
Allows the subscribers controlled access to the change data
June 17, 2014
Proprietary and Confidential
- 6 -
IGATE

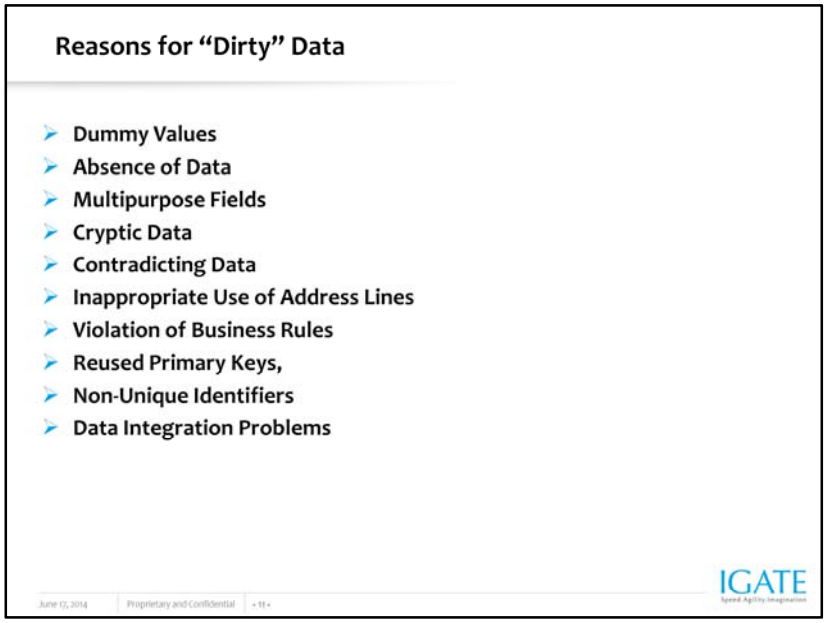

Change Data Capture
Subscriber
Subscriber needs to know what change data it is interested in
It creates a subscriber view to access the change data to which it has been granted access by the publisher
June 17, 2014
Proprietary and Confidential
IGATE

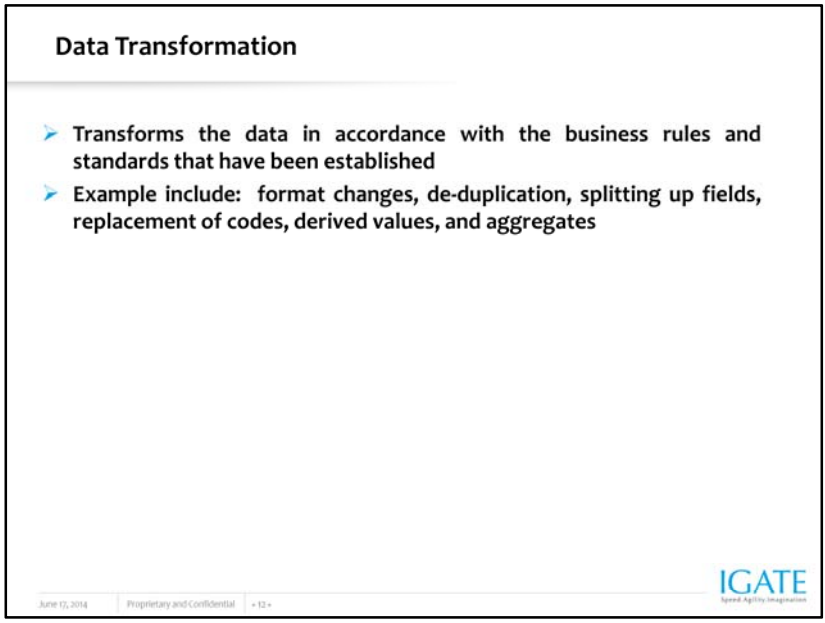


Data staging is used in cleansing, transforming, and integrating the data.

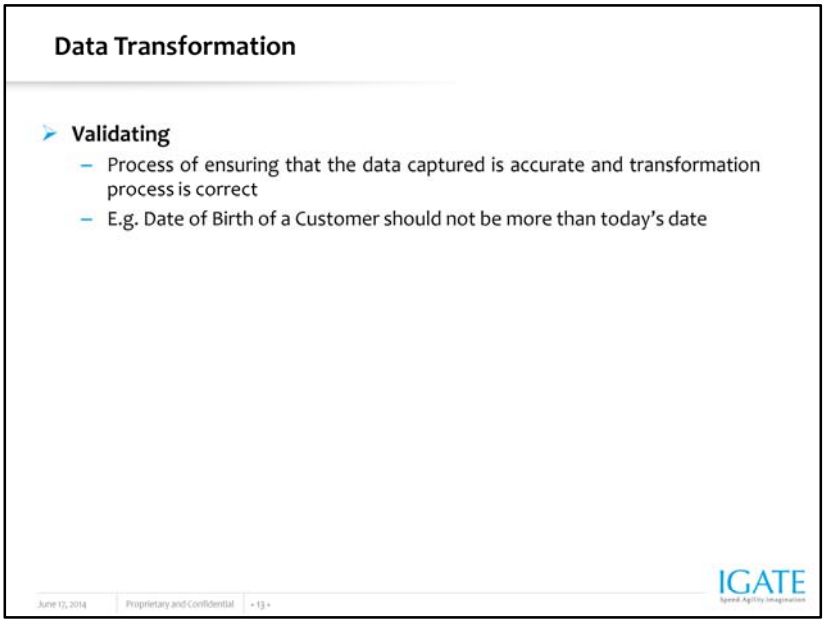

The ETL Process
Extract
Extract relevant data
Transform
Transform data to DW format
Build keys, etc.
Cleansing of data
Load
Load data into DW
Build aggregates, etc
June 17, 2014
Proprietary and Confidential
IGATE

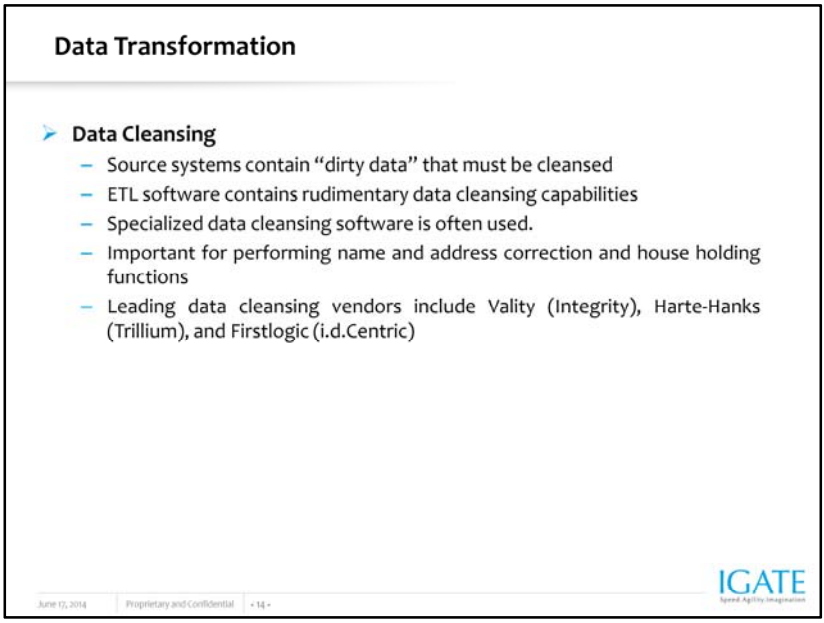

Data Extraction
Capture of data from Source Systems
Important to decide the frequency of Extraction
Sometimes source data is copied to the target database using the replication capabilities of standard RDBMS (not recommended because of "dirty data" in the source systems)
June 17, 2014
Proprietary and Confidential
« 10 »
IGATE


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
June 17, 2014
Proprietary and Confidential
IGATE

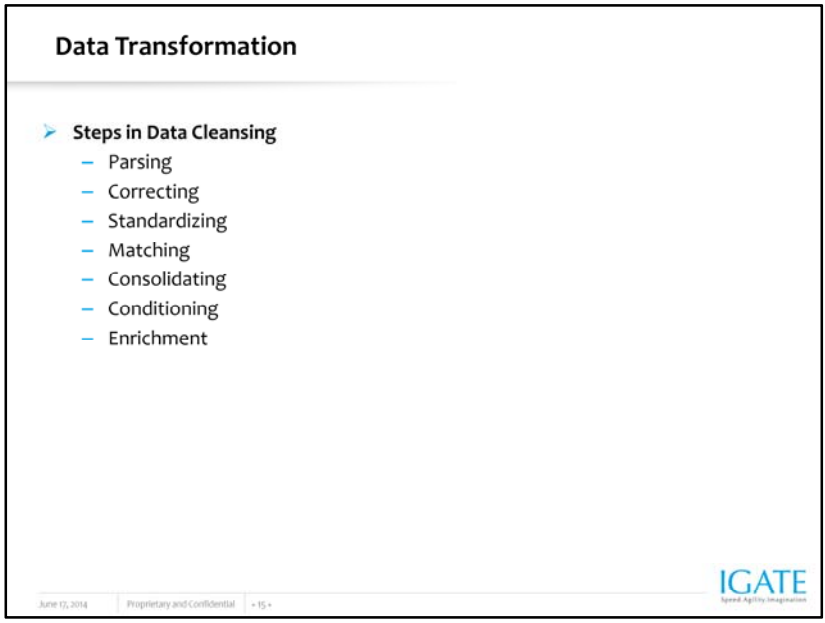


Aggregates, such as sales totals, are often precalculated and stored in the warehouse to speed queries that require summary totals.

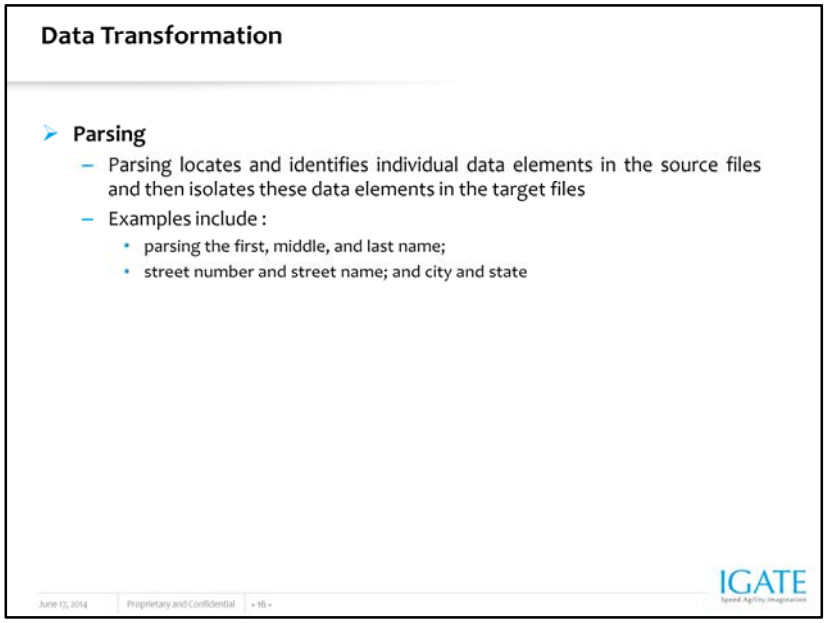

Data Transformation
Validating
– Process of ensuring that the data captured is accurate and transformation process is correct
– E.g. Date of Birth of a Customer should not be more than today's date
June 17, 2014
Proprietary and Confidential
« 13 »
IGATE



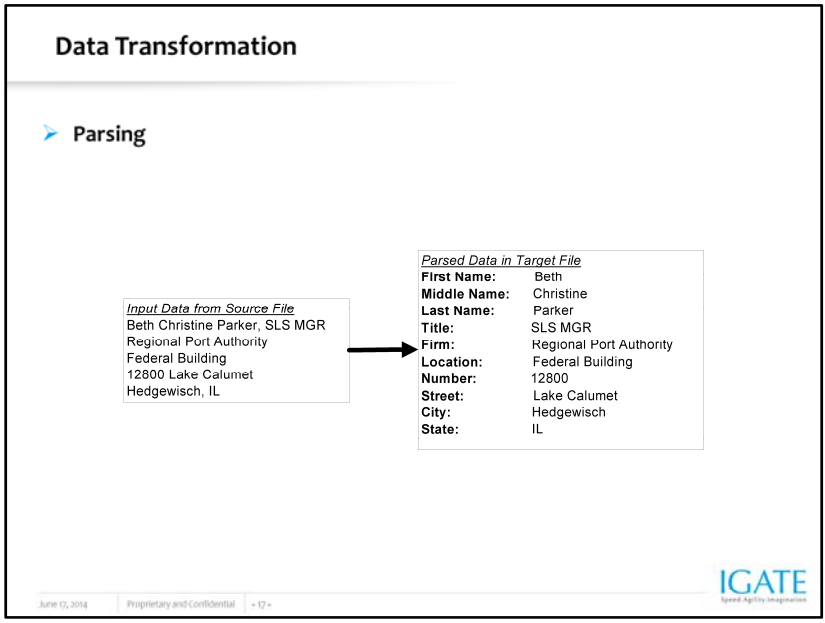
Data cleansing is critical to customer relationship management initiatives.

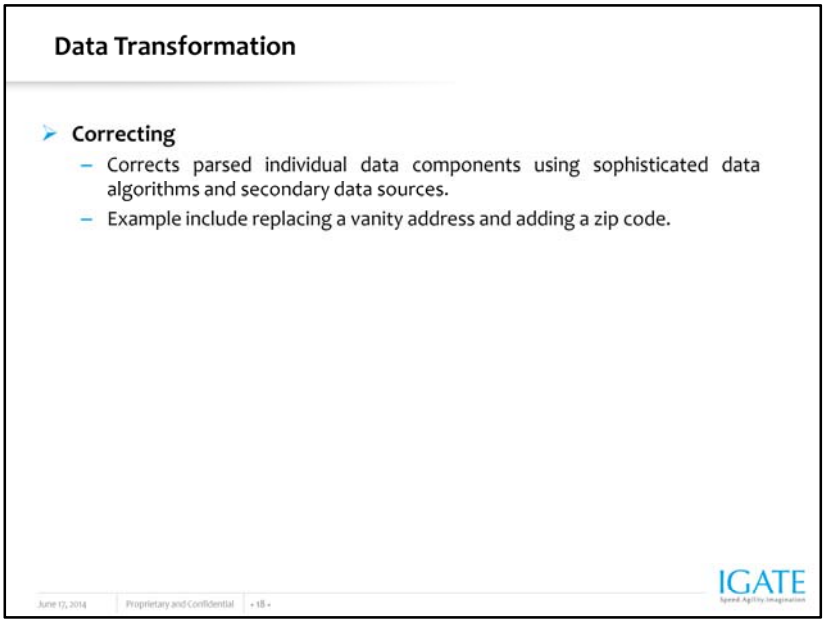


A good example to use is cleansing customer data. Most students can identify with receiving multiple copies of the same catalog because the company is not doing a good data cleansing job.

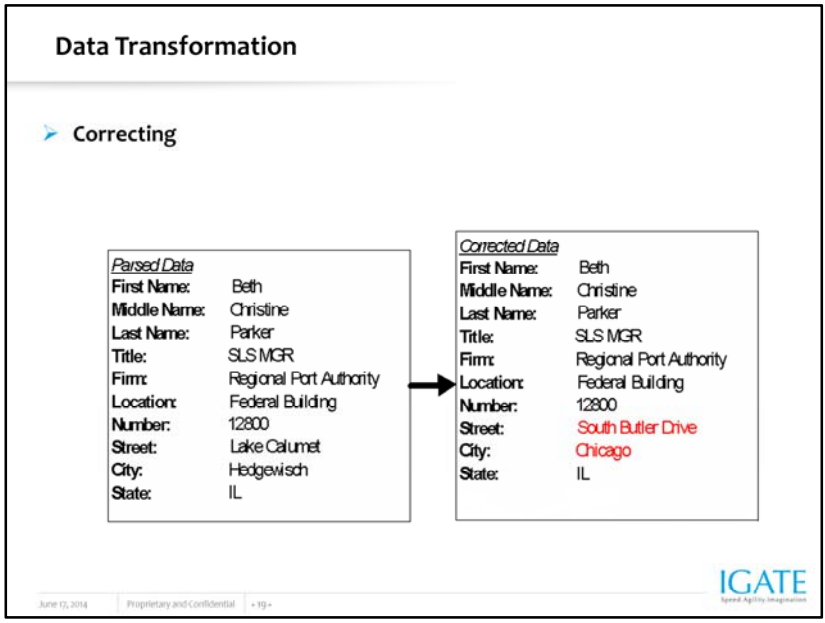


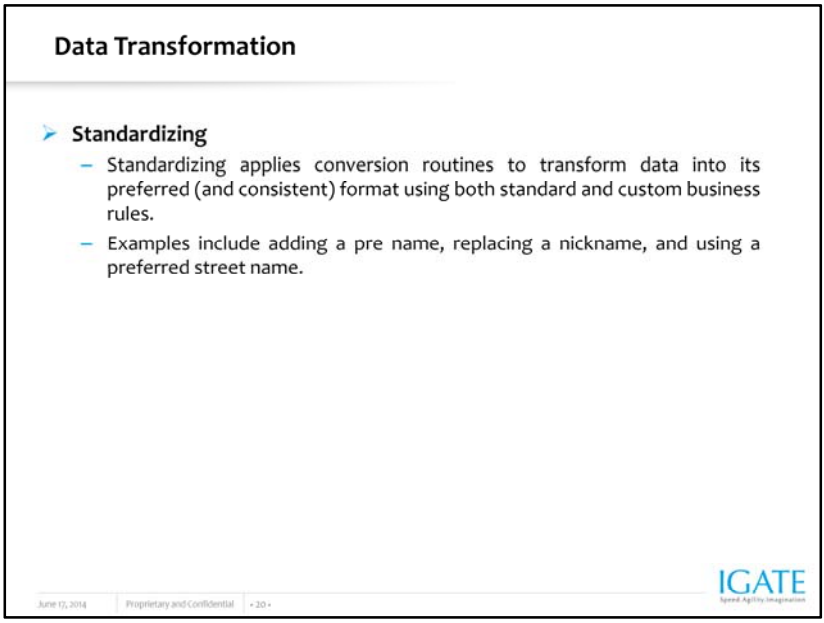
The record is broken down into atomic data elements.


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
June 17, 2014
Proprietary and Confidential
- 17 -
IGATE

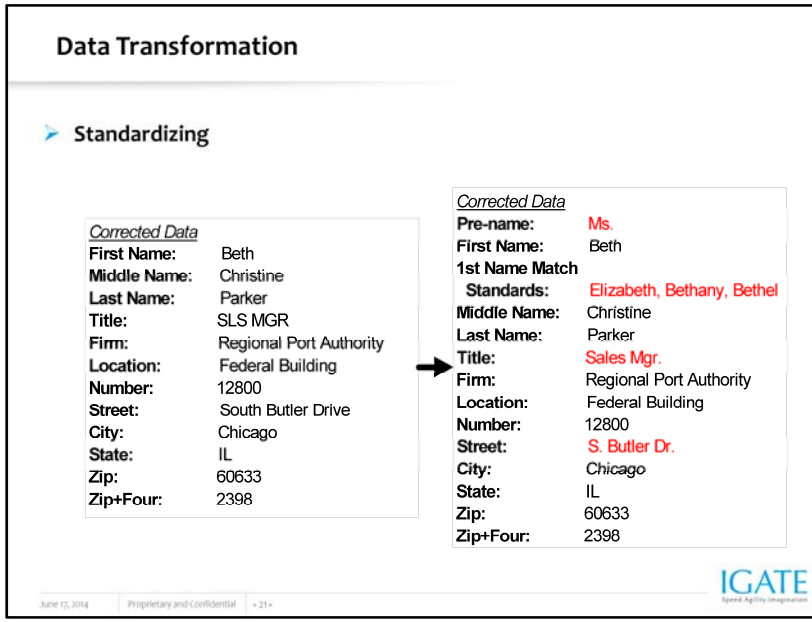


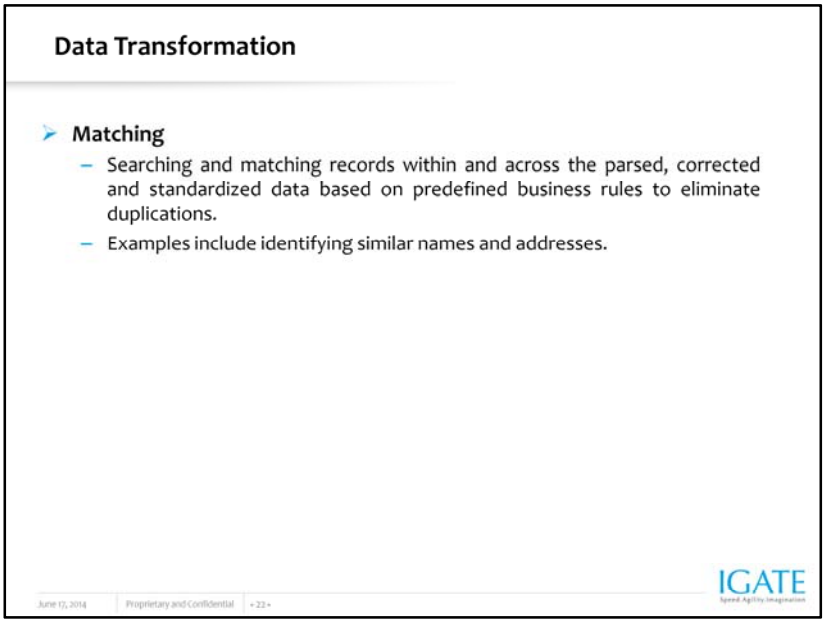
External data, such as census data, is often used in this process.


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
IGATE

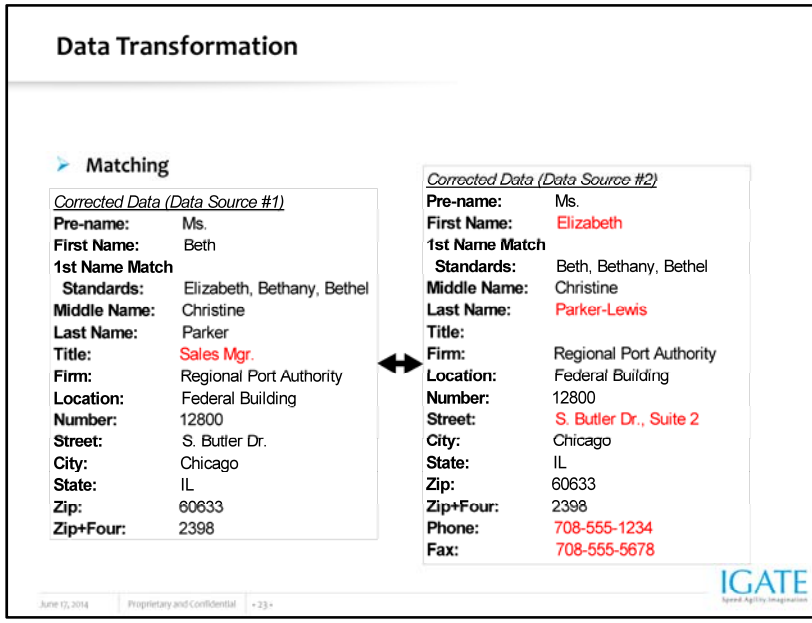


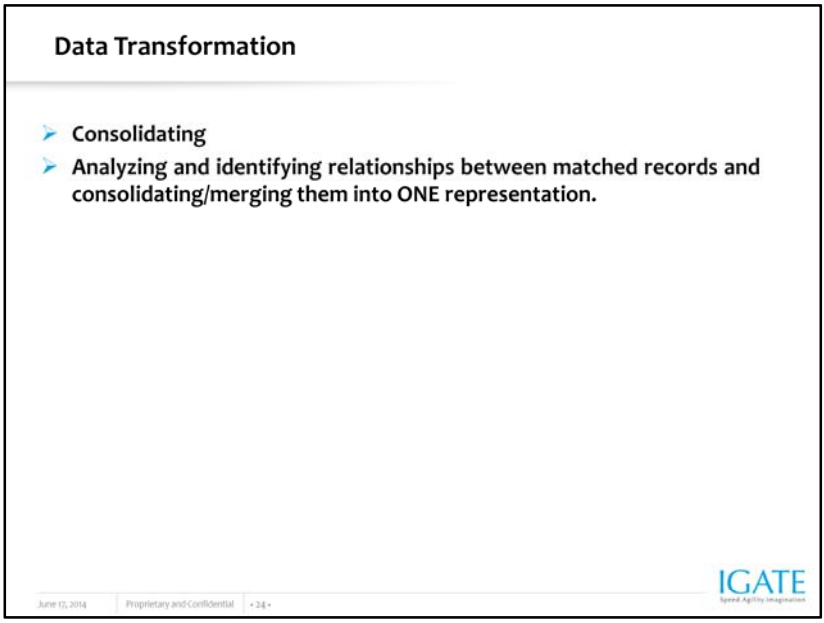
Companies decide on the standards that they want to use.


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
IGATE

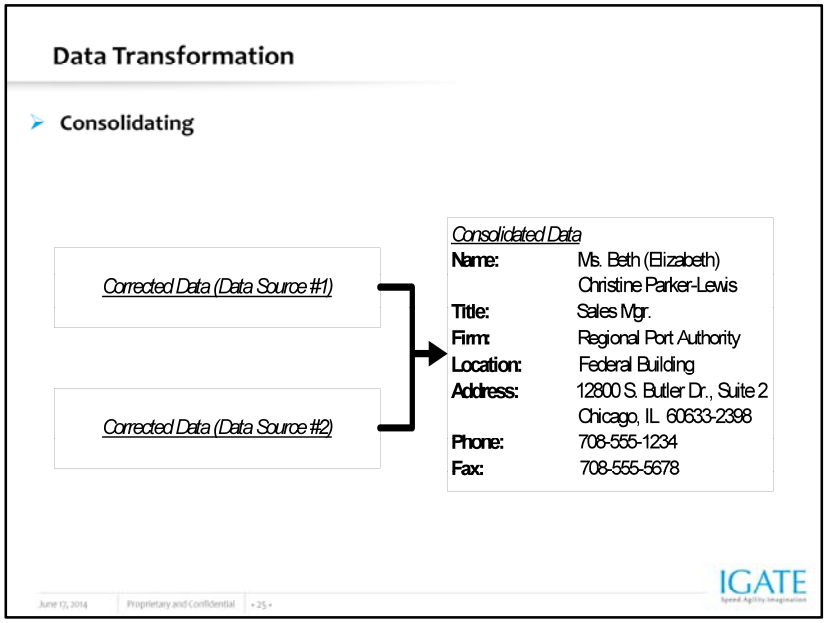


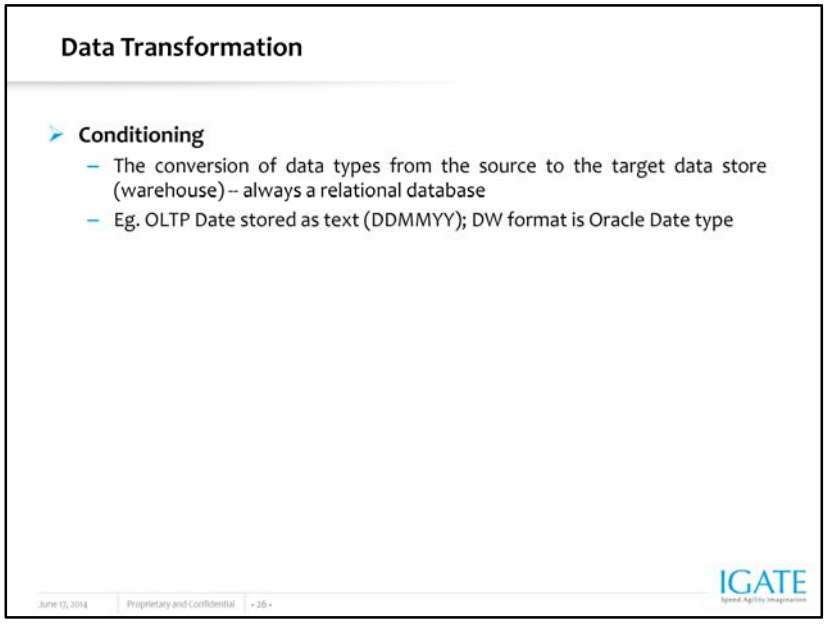
Commercial data cleansing software often uses AI techniques to match records.

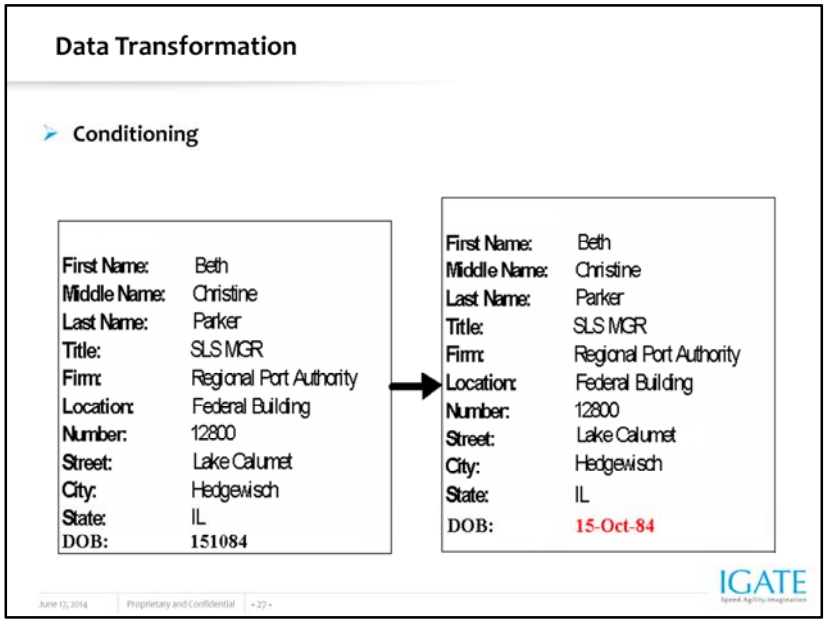

Data Transformation
Matching
Corrected Data (Data Source #1)
Pre-name: Ms.
First Name: Beth
1st Name Match Standards: Elizabeth, Bethany, Bethel
Middle Name: Christine
Last Name: Parker
Title: Sales Mgr.
Firm: Regional Port Authority
Location: Federal Building
Number: 12800
Street: S. Butler Dr.
City: 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
IGATE



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

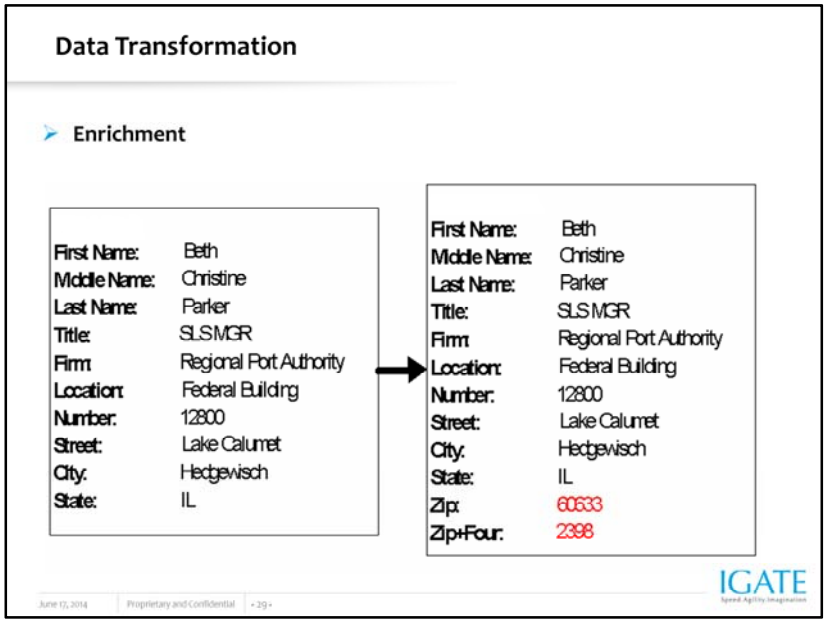

Data Transformation
Consolidating
Corrected Data (Data Source #1)
Corrected Data (Data Source #2)
Consolidated Data
Name: Ms. Beth (Elizabeth) Christine Parker-Lewis
Title: Sales Mgr.
Firm: Regional Port Authority
Location: Federal Building
Address: 12800 S. Butler Dr., Suite 2 Chicago, IL 60633-2398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

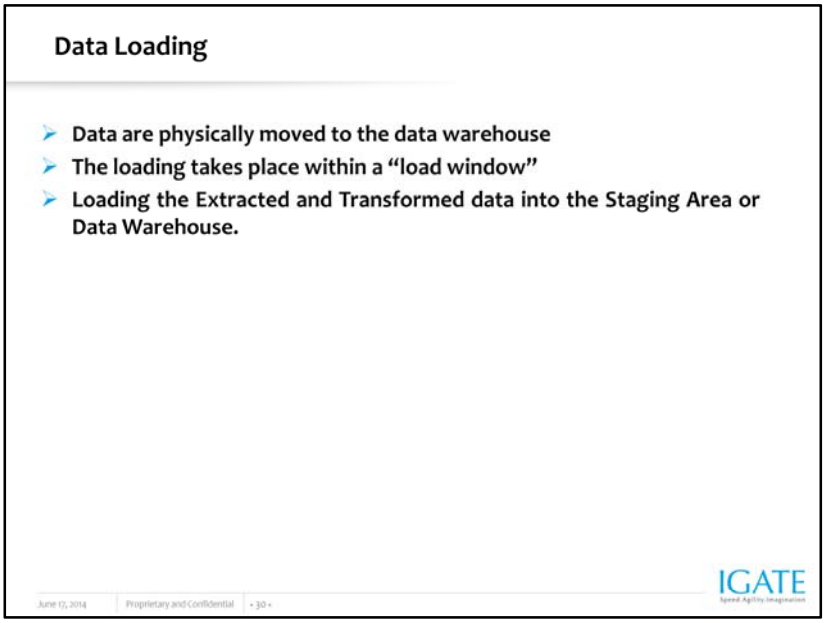

Data Transformation
Conditioning
– The conversion of data types from the source to the target data store (warehouse) -- always a relational database
– Eg. OLTP Date stored as text (DDMMYY); DW format is Oracle Date type
June 17, 2014
Proprietary and Confidential
- 26 -
IGATE


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
IGATE

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

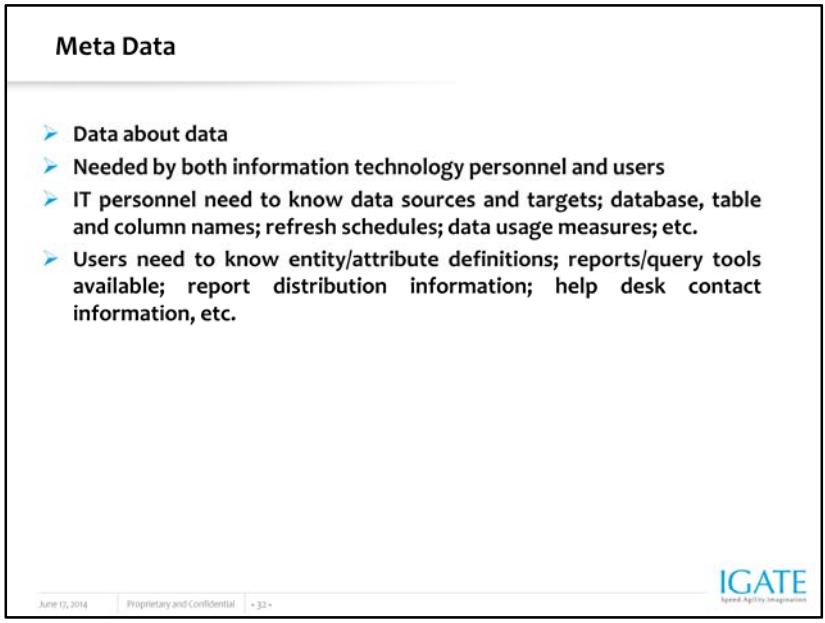

Data Transformation
Enrichment
First Name: Beth
Middle Name: Christine
Last Name: Parker
Title: SLS MGR
Firm: Regional Port Authority
Location: Federal Building
Number: 12800
Street: Lake Calumet
City: Hedgewisch
State: IL
First Name: Beth
Middle Name: Christine
Last Name: Parker
Title: SLS MGR
Firm: Regional Port Authority
Location: Federal Building
Number: 12800
Street: Lake Calumet
City: Hedgewisch
State: IL
Zip: 60633
Zip+Four: 2398
IGATE



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

Data Loading

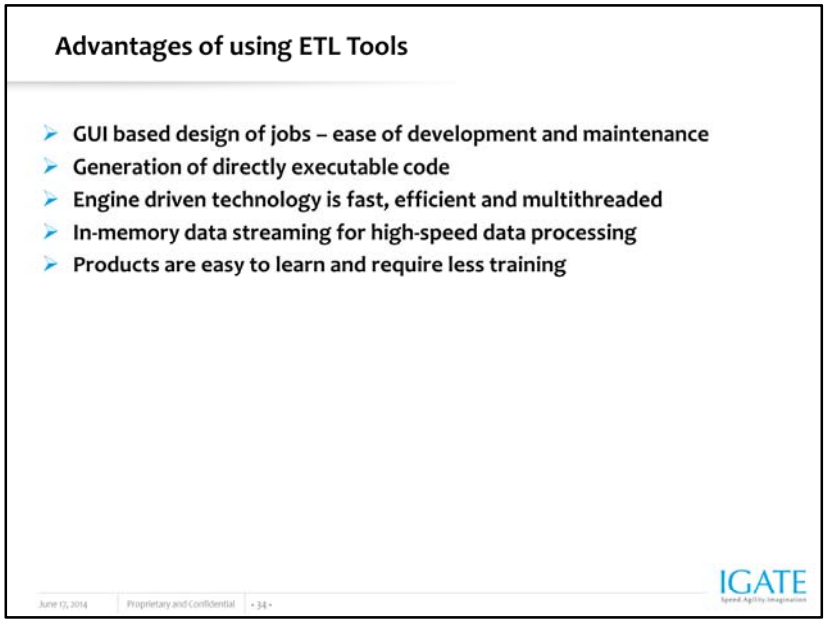
- First time bulk load to get the historical data into the Data Warehouse
- Periodic Incremental loads to bring in modified data
- The Loading window should be as small as possible
- Should be clubbed with strong Error Management process to capture the failures or rejections in the Loading process

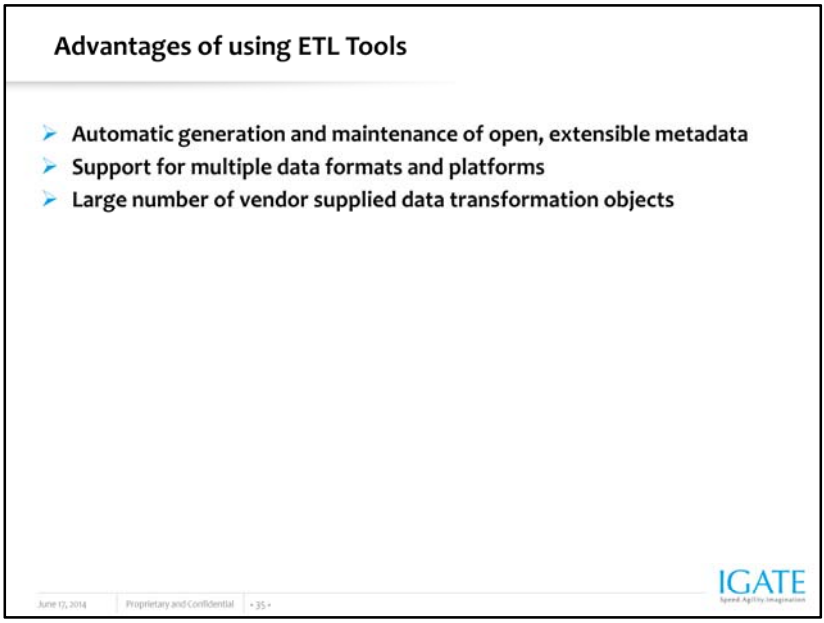


The importance of meta data is now realized, even though creating it is not glamorous work.

Feature of ETL Tools

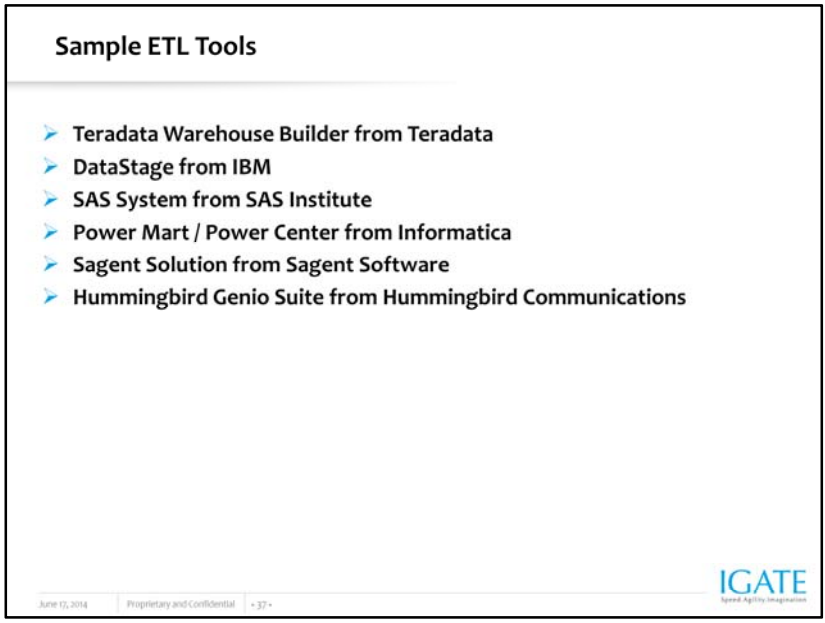
- Support data extraction, cleansing, aggregation, reorganization, transformation, and load operations
- Generate and maintain centralized metadata
- Filter data, convert codes, calculate derived values, map source data fields to target data fields
- Automatic generation of ETL programs
- Closely integrated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


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
June 17, 2014
Proprietary and Confidential
- 34 -
IGATE

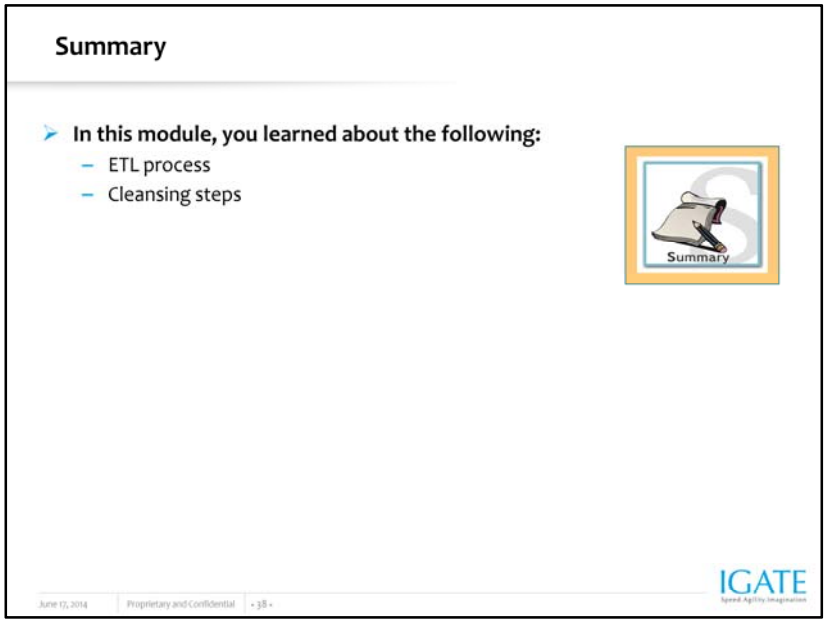

Advantages of using ETL Tools
Automatic generation and maintenance of open, extensible metadata
Support for multiple data formats and platforms
Large number of vendor supplied data transformation objects
June 17, 2014
Proprietary and Confidential
« 35 »
IGATE

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**



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



Add the notes here.