NANYANG THE INNOVATIVE POLYTECHNIC School of Information Technology

Data Cleaning and Wrangling

What we will cover

- Data ETL (Extract, Transform, Load)
- Data Format
- Data Import & Manipulation
- Data Cleaning
- Data Cleaning vs Data Wrangling

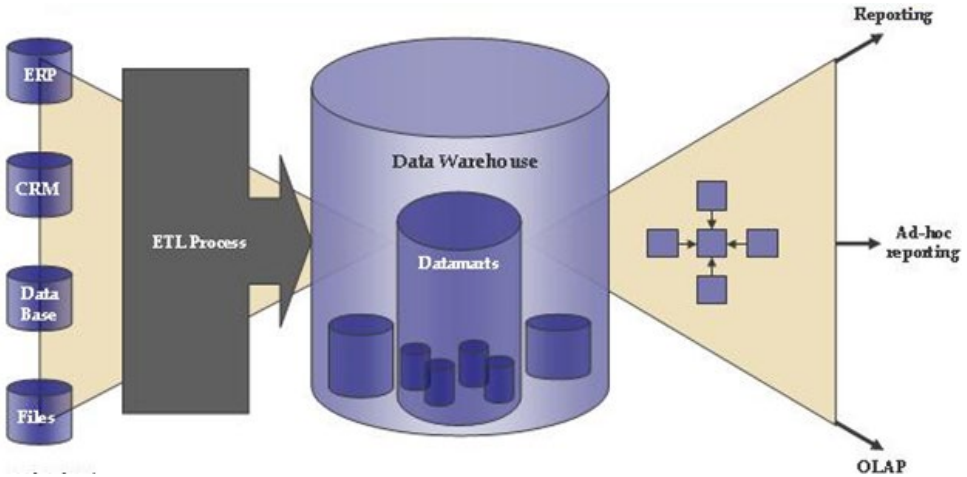
Data ETL

Data ETL

- Stands for Extract - Transform - Load
- Covers a process of how the data are loaded from the source system to the data warehouse
- Frequently encompass a cleaning step

Difference between data warehouse and data mart

Data warehouse	Data mart
enterprise-wide data	department-wide data
multiple subject areas	single subject area
difficult to build	easy to build
takes more time to build	less time to build
larger memory	limited memory



Purpose of Data ETL

ETL can be used:

- to acquire a temporary subset of data for reports or other purposes
- population of a data mart or data warehouse
- conversion from one database type to another
- migration of data from one database or platform to another

Data ETL

Extract

- Covers the data extraction from the source system and makes it accessible for transformation processing.
- Main objective is to read/retrieve all the required data from the source system with as little resources as possible.
- should be designed in a way that it does not negatively affect the source system in terms or performance, response time or any kind of locking.

Data ETL

Clean

- Cleaning step is one of the most important as it ensures the quality of the data in the data warehouse.
- Should perform basic data unification rules, such as:
  - Making identifiers unique (Eg: “Male”, “Man” to “M” and “Female”, ”Woman” to “F”)
  - Convert null values into standardized . (Eg: Not Available/Not Provided/0)
  - Convert phone numbers, ZIP codes to a standardized form
  - Validate address fields, convert them into proper naming, e.g. Street/St/ Str
  - Validate address fields against each other (State/Country, City/State, City/ZIP code, City/Street)
  - Joining together data from multiple sources (Eg: lookup, merge)

Data ETL

Transform

- The transform step applies a set of rules to transform the data from the source to the target . This includes converting any measured data to the same dimension (i.e. conformed dimension) using the same units so that they can later be joined.
- The transformation step also requires joining data from several sources, generating aggregates, sorting, deriving new calculated values, and applying advanced validation rules.

Data ETL

Load

- To write the resulting data (either all of the subset or just the changes) to a target database, which may or may not previously exist.

Managing the ETL Process

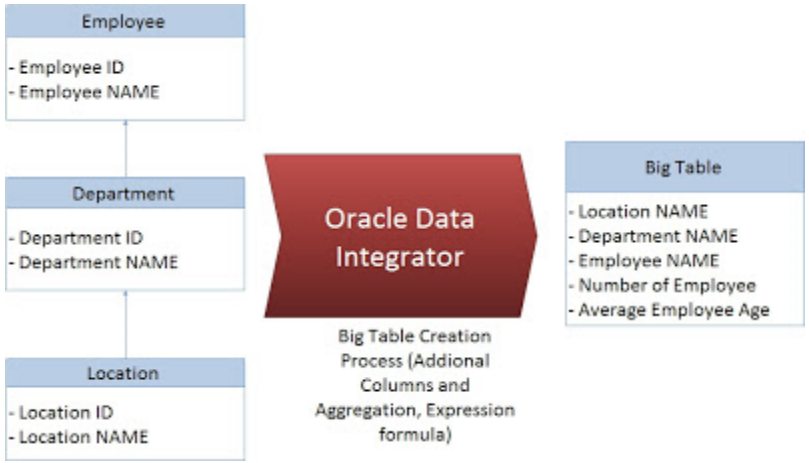
As with every application, there is a possibility that the ETL process fails. This can be caused by missing extracts from one of the systems, missing values in one of the reference tables, or simply a connection or power outage. Therefore, it is necessary to design the ETL process keeping fail - recovery in mind .

Staging

- To store intermediate results of processing before entering data warehosue. The staging area should is be accessed by the load ETL process only.

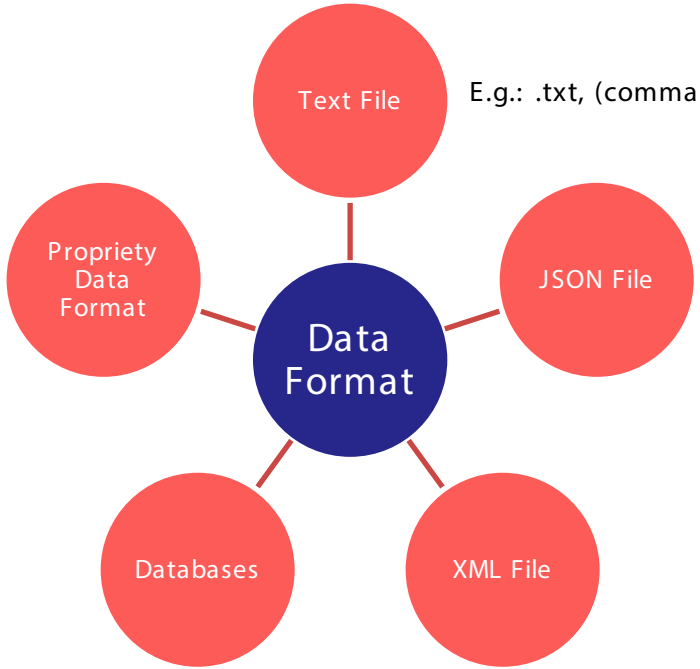
Data ETL Tools

- Oracle Data Integrator
- SAP Data Integrator
- SQL Server Integration Services (SSIS)



Data Format

Different Types of Data Format



E.g.: .txt, (comma delimited) .csv, .tab

E.g.: IBM SPSS File Format

Text File

- .txt
- .csv

```
Anna Smith,42,female,korea

Peter Jones,35,male,Singapore
```

- .tab (or . tsv)

```
Anna Smith    42        female      korea

Peter Jones   35        male        singapore
```

XML File

- XML stands for e**X**tensible **M**arkup **L**anguage

```
<employees>
  <employee>
    <firstName>Anna</firstName>
    <lastName>Smith</lastName>
  </employee>
  <employee>
    <firstName>Peter</firstName>
    <lastName>Jones</lastName>
  </employee>
</employees>
```

JSON File

- JSON stands for JavaScript Object Notation

```
{
"employees":
 [
   { "firstName":"Anna" , "lastName":"Smith" },
   { "firstName":"Peter" , "lastName":"Jones" }
 ]
}
```

Handling text File

- Programming Language, e.g. Python, Java
- MS Excel
- OpenRefine
  http://openrefine.org/

  - OpenRefine (formerly Google Refine) is a powerful tool for working with messy data. It is a standalone open source desktop application
  - For data cleanup; transforming it from one format into another; it behaves more like a database.
  - It operates on *rows* of data which have cells under *columns,* which is very similar to relational database tables and extending it with web services and external data.

Data Import and Manipulation

Data Import

- Automated or semi - automated input of data between different software.
- Involves "translating" from the format used in one application into that used by another, where such translation is accomplished automatically via machine processes, such as data transformation, and others.

Data Import

- Any statistical analysis presumes that you have the appropriate data in a format suitable for analysis.
- The data might exist in different format:
  - In an ***Excel® file*** – which might still need to be rearranged to get it in the form of a rectangular data set.
  - In a ***text file*** – which is any file that can be opened and read in a text editor such as Notepad; it can be imported into Excel using Excel's text import wizard.
  - In a ***relational database*** (such as Access, SQL Server, Oracle)—which can be imported into Excel by forming a query using the Microsoft Query package.
    - A query specifies exactly which data you want to import.
- Once the data is imported, it may need to be cleansed to fix wrong values.

Data Manipulation

- Process of taking data and manipulating it in a method to be easier read or organized.
- For example, a log of data entries could be organized in alphabetical order, making it easier to view (grouping) and find information (indexing).

Data Cleaning

Data Cleaning

- Data can be stored in subsets or various formats so conversion or extraction is needed before visualisation can be take place

- Data may need to be cleaned
  - Any missing values? Duplicate data?
  - Smooth noisy data, identify outliers

Missing Values

- Missing value != error in data
- Reasons for missing values
  - Information is not collected
    (e.g., people decline to give their age and weight)
  - Attributes may not be applicable to all cases
    (e.g., annual income is not applicable to children)
  - Poor data capturing form /database design
    (e.g., required or optional field)
- Handling missing values
  - Eliminate Data Objects
  - Ignore the Missing Values During Analysis
  - Estimate Missing Values
  - Replace with All Possible Values

ID	Income
1	3890
2	
3	3243
4	3021
5	3452

Handling Missing Values

1. Ignore the Missing Values During Analysis
    - Used when field is not labelled
    - Used when the row of data contains several fields with missing values
    - Can result in poor data
    - Typically, we will ignore the records if it's less than 5%

2. Estimate Missing Values
    - Time-consuming
    - May not be feasible if given a large data set

Handling Missing Values

3. Replace With All Possible Values – Use a global constant
    - Replace with the same value e.g. Unknown, -∞

4. Replace With All Possible Values – Use the attribute mean or median for all samples belonging to the same class as the given record
    - E.g. record with missing value is a customer with credit score of 3. Use the average income of $60,000 for all customers with credit score of 3.

Handling Missing Values

5. Replace With All Possible Values – Use the most probable value
    - Use methods like regression, inference based tools to determine a likely value
    - Popular strategy – uses the most information from present data to estimate the missing value

Duplicate Data

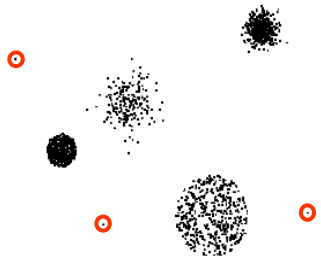
- Data set may include data objects that are duplicates, or almost duplicates of one another
  - Major issue when merging data from different sources
- Examples:
  - Same person with multiple email addresses
  - Amazon products, Qoo10 products

ID	Email address
232	rainbow.chang@yahoo.com
345	changyang@gmail.com
455	sherry797@gmail.com
455	sherry.yang@hotmail.com
678	jennye@yahoo.com.sg

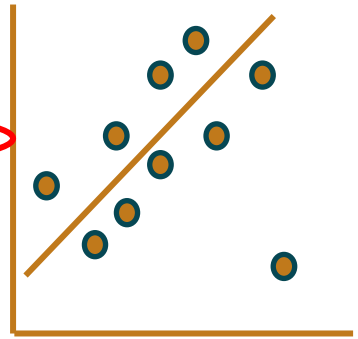
- Data cleaning
  - Duplicate rows are a common problem.
  - Filter for unique values first to confirm that the results are what you want before you remove duplicate values.

Outliers

- Outliers are data objects with characteristics that are considerably different than most of the other data objects in the data set
- When to include outliers?
  - fraud cases
  - major trends



ID	Income
1	3890
2	156
3	3243
4	3021
5	3452



Noisy Data

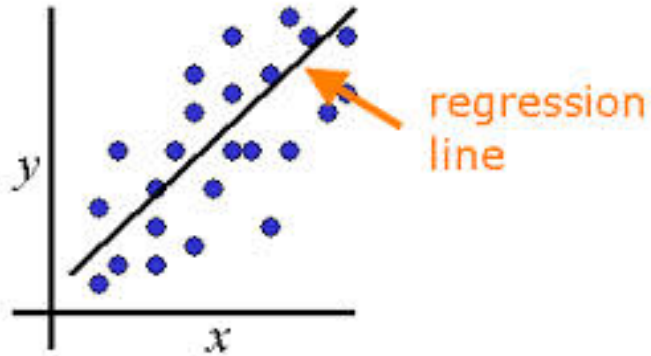
- What is Noise?
  - Noise is random error or variance in a measure variable

- “Smooth” out the data to remove noise

Smoothing Noisy Data

Regression

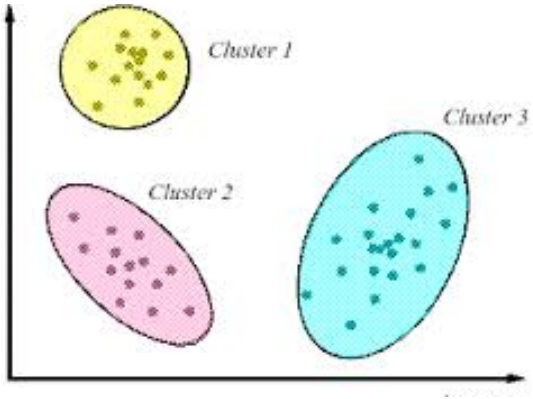
- Smooth data by fitting the data to a function
- E.g. Linear Regression – find “best fit” line to fit 2 attributes



Smoothing Noisy Data

Clustering

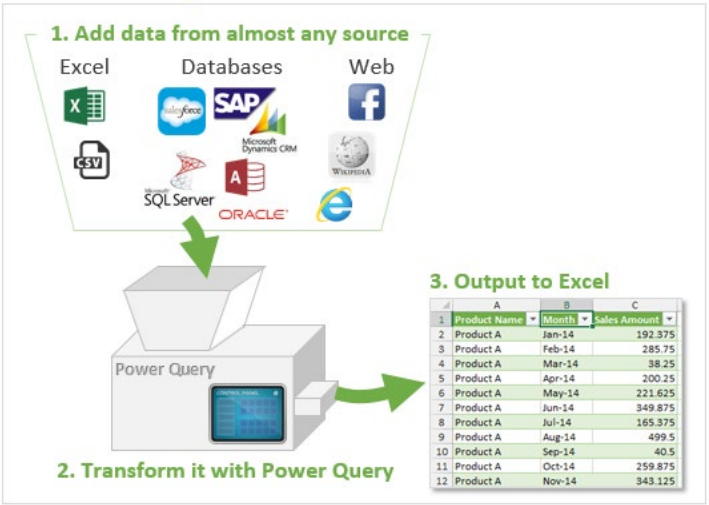
- Similar values are organised into clusters (groups)
- Values that are outside of the clusters are considered outliers that should be removed



Data Transformation

These data transformations could include tasks like:

- Remove columns, rows, blanks
- Convert data types – text, numbers, dates
- Split or merge columns
- Sort & filter columns
- Add calculated columns
- Aggregate or summarize data
- Find & replace text



Data Cleaning vs Data Wrangling

Data Cleaning vs Data Wrangling

Data Cleaning (aka Data Cleansing)

- Process of finding and correcting inaccurate data
- Main objective → to identify and remove inconsistencies without removing data in order to perform data analysis
- Type of Data Cleaning Functions
  - Find and Replace
  - Add missing values
  - Remove duplicate rows
  - Spelling checks
  - Formatting data

Data Wrangling (aka Data Munging)

- Focuses on transforming data's format by converting "raw" data into another format more suitable for use
- Sometimes data cleansing is part of data wrangling process
- Types of Data Wrangling Functions
  - Joins
  - Filters
  - Group By functions
  - Merging
  - Restructuring data (normalising, etc)