Managing Data & Databases

Session 13
Data Extraction, Transformation and Loading (ETL)

The ETL Process

- Many tools
  - Direct import + ad-hoc transformation using database functions
  - Custom-made applications
  - Off-the-shelf solutions
    - Talend, Clover, etc.

- Cleansing: ECTL is a better acronym
  - Database functions
  - Custom-made applications
  - Ready-made solutions
    - OpenRefine, etc.

Today's Dose of SQL

- DML
  - String Functions
  - String Comparison
    - Wildcards
    - Regexp
  - Casting Functions
  - Date and Time Functions

String Functions

- *Use case:* Matching, comparing and transforming strings
  - Substring
  - Locate
  - Like
  - Trim
  - Ltrim
  - Rtrim
  - Regexp
  - Length

Ref: https://dev.mysql.com/doc/refman/5.5/en/string-functions.html

String Comparison

- Two main methods in MySQL
  - Using LIKE and a wildcard
    - _ matches exactly one character
    - % matches zero or more characters
  - Using REGEXP and a regular expression
    - Regular expressions almost constitute a programming language for text matching
    - To learn: http://www.regular-expressions.info
    - To test: http://www.regexplanet.com

Casting and conversion

- *Use case:* Converting variables from one data type to another
  - Cast
  - Convert

Ref: https://dev.mysql.com/doc/refman/5.5/en/cast-functions.html

Date and Time Functions

- *Use case:* Comparing, converting or doing basic operations on dates, times and intervals
  - Curtime
  - Curdate
  - Now
  - Second
  - Timediff
  - Datediff

Ref: https://dev.mysql.com/doc/refman/5.5/en/string-functions.html

Two Types of Join

- SELECT t1.a, t2.b
  FROM t1
  JOIN t2
  ON t1.c = t2.d

- SELECT t1.a, t2.b
  FROM t1,t2
  WHERE t1.c = t2.d

- Update joins are of the second kind