University of California, Berkeley
Master of Information and Data Science (MIDS)
W205 – Fundamentals of Data Engineering

Week 7 – Data Wrangling, Part II: ETL (Extract, Transform, Load), ELT, and Data Cleansing

Agenda for Today's Class

- Attendance and Participation
- Announcements
- Schedule and Due Dates
- Work / Life / School Balance
- Asynch High Level Review in a Nutshell
- Breakouts
- Summary

Attendance and Participation

Please record your attendance and participation for today's class:

GitHub => ucb_mids_w205_repo => README.md => Attendance and Participation

Announcements

- Upcoming holidays and/or breaks
- Makeup classes for holidays
- Upcoming events
- Student evaluations
- Etc.

Schedule and Due Dates

Take a quick look at the next couple of weeks' due dates:

GitHub => ucb_mids_w205_repo => README.md => Schedule and Due Dates

Work / Life / School Balance
Open Discussion

Student feedback

- About 5 minutes
- How are things going related to work / life / school balance?
- How is w205 going? Difficulty? Time?
- Impact of any natural and/or man-made disasters
- Etc.

Asynch High Level Review in a Nutshell

Each week we will spend about 15 minutes reviewing the most important high level concepts from the asynch

Waterfall versus CI/CD

- Waterfall
  - Each step must be completed in its entirety before next step
  - No going back - stuck with bad decisions for years
  - Takes months or years to see any results
  - Consulting companies / outsourcing still use to maximize billing
- CI / CD
  - Continually Integrating / Continually Developing
  - Gaps can be billed at any time
  - Agile – able to turn on a dime as our business changes

ETL / ETL

- ETL
  - Extract Transform Load
  - Often ETL is used now to mean or include ELT
  - Old school style pure ETL is obsolete
- ELT
  - Extract Load Transform
  - Load data as early in the pipeline as possible to take advantage of tools
- ETL / ELT are part of Data Wrangling
  - Data Wrangling is a lot more than simple transformation

Staging Tables

- Load raw data into staging tables as soon as possible
  - Data exploration
    - SQL, functional programming, tools, tabular data charts, data visualizations
  - Data transformation
  - Data cleansing

Data Cleansing

- Typing and Spelling Issues
  - Typos, misspellings, alternate spellings, garbled data
- Soundex
  - Phonetic algorithm
  - Pre-dates computers
- Levenshtein Distances
  - Number of additions, deletions, and changes between 2 strings
  - Can be used as distance measure in string kernels in ML
- Fuzzy Logic
  - Traditional algorithms, plus context, grammar, statistics, etc.

Data Cleansing (continued)

- Dedupe
  - Remove duplicates
  - Danger - sometime legitimate
- Missing Values
  - Imputing – filling in missing values
  - Average, fill down, null value, predictive model, etc.
  - Danger – always a risk we will distort data
- Outliers
  - Keep or replace?
  - Keep if actual exception, replace if bad data

Data Cleansing (continued)

- Validation Rules
  - String in date, invalid date, string in number, number in string, overflow on field length, etc.
- Lookup Tables
  - Value not present in lookup table, human error
- Referential Integrity
  - Similar to lookup tables, but programming error, not human error
- Consistency, Contradictions
  - One piece of data says one thing, another piece contradicts it

Data Cleansing (continued)

- Completeness
  - One piece of data implies another piece of data should exist and it does not
- Uniformity
  - Data the same
  - Units of measure
  - What’s included in a number

Breakouts

GitHub => ucb_mids_w205_repo => breakouts

(time permitting, we may not get to all of them)

Summary

Instructor will give a brief (about 2 minute) summary of today’s class.