Data Wrangling

Concepts

Data Wrangling

- Data engineers prepare data for data science
- Analogous to cattle wrangling from the wild into pens
  - Cattle out on the range
  - Find them in several groups
  - Herd them into one group
  - Move them into pens
  - Clean up, vet, etc.

Munging

- Mung—programmer slang—changes to a piece of data or file that are destructive and irrevocable
- Derogatory term originally
- Data wrangling probably better term
- Some say data munging is part of data wrangling

Data Encoding

- Bit, nibbles, bytes
- Storage formats
  - Integers
  - Floats
  - Characters
  - International characters

File Formats

- CSV (comma-separated values)
- JSON (JavaScript Object Notation)
- Excel workbooks and worksheets
- Compressed archives: ZIP, 7-Zip, Linux tarballs

ETL (ELT): Extract, Transform, Load

- E**T**L
  - Popular with traditional data warehousing
- E**L**T
  - Popular with big data

ETL (ELT): Extract, Transform, Load (cont.)

- Staging tables
- Raw data exploration using staging tables
- Transforming data
  - Parsing
  - Joining
  - Augmenting
  - Consolidation
  - Filtering

Data Cleansing

- Typos, misspellings, etc.
  - Fuzzy logic, Soundex, Levenshtein distances
- Dedup (de-duplicating): removing duplicates
- Imputing: filling in missing values
- Outliers: throw away or keep
- Validation rules
- Lookup tables

Data Cleansing (cont.)

- Referential integrity: child rows match to a parent row
- Consistency
- Contractions
- Completeness
- Uniformity

The End

Data Wrangling

Business Cases

Third Party Sales Channels

- Receive sales data from third party sales channels
- Validate, clean, merge, and load into our analytical sales database

Receive Data

All vendors have a different way to send us data.

- Live streams
- Download files from website
- Email us files as attachments

File Formats

All vendors have a different file format to send us data.

- CSV
- JSON
- Excel workbooks

Validation Varies

- Validation varies per vendor
- Unvalidated manual data entry for some vendors—very error prone
- Some vendors do not validate against our customer data that we give them
- Problems with consistency, contradictions from some vendors
- Missing values, missing rows, etc. from some vendors

Data Wrangling

Obviously, we have a lot of data wrangling to do.

- Several ways to pull in data
- Several file formats
- Must explore and validate everything
- Must clean everything to the point where it’s safe to load into our sales analytical database

The End

Building Blocks of Storage and Encoding

Concepts

Bit

- Binary digit
- Base 2
- 0 or 1

Nibble (Nybble)

- 4 bits
- 1 hex digit
- Half byte

Byte

- 8 bits
- 2 hex digit (nibble pair)

Kilobyte

- 1024 bytes
- KiB
  - Official abbreviation
  - Guarantees 1024
- KB or K
  - Could be 1024?
  - Could be 1000?

Megabyte

- 1 MiB = 1024 KiB = 1024^2 bytes
- MiB
  - Official abbreviation
  - Guarantees 1024^2
- MB or M
  - Could be 1024^2?
  - Could be 1024 * 1000?

Gigabyte

- 1 GiB = 1024 MiB = 1024^3 bytes
- GiB
  - Official abbreviation
  - Guarantees 1024^3
- GB or G
  - Could be 1024^3?
  - Could be 1024 * 1000 * 1000?

Terabyte, Petabyte, Exabyte

- Terabyte
  - 1 TiB = 1024 GiB = 1024^4 bytes
- Petabyte
  - 1 PiB = 1024 TiB = 1024^5 bytes
- Exabyte
  - 1 EiB = 1024 PiB = 1024^6 bytes

Zettabyte, Yottabyte

- Zettabyte
  - 1 ZiB = 1024 EiB = 1024^7 bytes
- Yottabyte
  - 1 YiB = 1024 ZiB = 1024^8 bytes

Hexadecimal

- Base 16
- 0 through 15 inclusive
- 0, 1, 2, 3, 4, 5, 6, 7, 8, 9,
  A = 10, B = 11, C = 12, D = 13, E = 14, F = 15
- 1 hex digit = 4 bits
- 1 hex digit = 8 bits = 1 byte
- Often express binary data in hex with 2 hex digits for each byte

Encoding English

- EBCDIC
  - Extended Binary Coded Decimal Interchange Code
  - IBM mainframes
- ASCII
  - American Standard Code for Information Exchange
  - 7 bits for English: uppercase, lowercase, numbers, punctuation

Unicode

- Languages that use the Latin alphabet may have a few extra characters.
- Other languages may have 2 or 3 bytes needed to represent their language.
- Problem: If most of our data is in English and we convert everything to Unicode, we double or triple our database size.

UTF-8

- 8-bit Unicode Transformation Format
- Best of both worlds
- ASCII needs 7 bits; a byte has 8 bits, so 1 extra bit
- No wasted space
  - If the extra bit is off, one byte for the character
  - If the extra bit is on, multiple bytes for the character
- Emojis are part of Unicode and handled by UTF-8
  - Allows for emoji analytics since they often change the meaning of a statement

Uuencoding

- Invented by Mary Ann Horton, UC Berkeley, 1980
- Allows binary data to be encoded into regular characters so it can pass through networks
- Started with email, used for internet, web pages, downloads, etc.
- MIME (multipurpose internet mail extensions)
  - Base64—newer version of uuencoding

The End

Building Blocks of Storage and Encoding

Business Cases

Consumer Ratings Website

- Consumers can rank businesses with 1 to 5 stars and put in comments.
- Consumers want to use emojis in their comments.
- Consumers might also want to use languages other than English for their comments.
- Consumers can post pictures as well.

Solution

- Use UTF-8 to store data
  - English will use 1 byte per character with no wasted space
  - Supports emojis
  - Supports all languages with multiple bytes as needed
- Use MIME with base 16 encoding for pictures
  - Allows pictures to pass through the public internet with no binary format issues

Specify Storage Precisely

- We want to specify storage precisely
- We understand there are variations in sizing for megabytes, gigabytes, terabytes, etc.
- Solution
  - Specify everything in official notation, such as KiB, MiB, GiB, TiB, etc. so there are no misunderstandings

The End

CSV (Comma-Separated Values)

Concepts

CSV: Comma-Separated Values

- Oddly, there is no official standard for CSV.
- MS (Microsoft) Office's default for CSV has become the de facto standard in the absence of an official standard.

CSV Format

- Mimics the structure as a table in a relational database
- First line
  - Optional but usually present
  - List of field (column) names separated by commas
- Remaining lines
  - One line per record (row) with each field (column) separated by commas

Exceptions

- What if a field (column) has a comma in it?
  - We will not be able to tell if the comma represents the end of the field or a comma in the data
  - Solution: enclose the field (column) with double quotes (“ ”)
- What if a field (column) is enclosed in double quotes and it has a double quote in it?
  - Solution: two double quotes in sequence

Relational Tables

- CSV mimics the structure of a relational table
- Easy to load data from CSV into a relational table
- Easy to dump data from a relational table into CSV format
- Some products can even do SQL against CSV files, treating them like relational tables
  - Serverless SQL: will cover later this semester

The End

CSV (Comma-Separated Values)

Business Cases

Data Downloads

- CSV file(s) are very common
  - ZIP file for multiple CSV files
- Each CSV file mimics a database table
  - Easy to create a set of database tables and load in the data
    - Some datasets come with the SQL DDL for the tables and load scripts
  - Issue: foreign keys

In-House Products

- Most in-house products have import/export in CSV format
- A lot of knowledge in products outside of the IT department
- Want to tap that knowledge for data science: AI, ML, DL, etc.

The End

JSON (JavaScript Object Notation)

Concepts

XML: Extended Markup Language

- Predecessor to JSON
- Human readable
- Computer readable
- Markup
  - Start tag, end tab, attributes
- Content
  - Elements

JSON: JavaScript Object Notation

- Lighter weight version of XML
- Key: value pairs
- Lists
- Very similar to Python dictionaries and lists
- Nest multiple levels deep

Flat JSON File

- List of dictionaries
- Each dictionary has the same keys in the same order
- Works just like CSV
- Easy to load into a single database table
- Easy to dump a single database table into a flat JSON file

Nested JSON File

List of dictionaries

- A key’s value is a list of dictionaries
  - A key’s value is a list of dictionaries
    - Etc.

Nested JSON File and Relational Tables

- Top level JSON would be equivalent to a relational table
- Each nested list would have to be a separate relational table
- Advantage
  - JSON can hold data from several tables
  - NoSQL document databases are based on this concept
- Disadvantage
  - Loading and dumping to and from relational tables involves multiple tables and complicated primary key logic to work

Holes in JSON

- Two dictionaries at the same level of nesting have different keys
- Solution
  - Make a list of all possible keys
  - Default values (or null) for missing keys
- Issue
  - Few products handle this well
  - A lot of programming needed to check and fill in holes

The End

JSON (JavaScript Object Notation)

Business Cases

Data Downloads

- JSON as an alternative to CSV
- We want everything in one file
- We are going to load into other than a relational database
  - NoSQL database, especially
  - Data structures in memory
- We are going to load into a relational database
  - Tools which can infer JSON structure and create and load tables for us

In-House Products

- Most in-house products have import/export in JSON in addition to CSV
  - Same reasons as previous slide
- Some products may only have JSON
- A lot of knowledge in products outside of the IT department
- Want to tap that knowledge for data science: AI, ML, DL, etc.

Other Common Uses for JSON

- Enterprise message queues
  - Publisher-subscriber
  - Producer-consumer
  - Streaming data
- Web APIs
- Will cover more later this semester

The End

Excel Workbooks and Worksheets

Concepts

MS Excel

- MS Excel widely used at companies
- A lot of knowledge in MS Excel outside of the IT department
- Want to tap that knowledge for data science: AI, ML, DL, etc.
- A lot of users want the output of data science to be provided in MS Excel format—that is what they are comfortable using

MS Excel Workbooks

- Collection of MS Excel worksheets
- Each worksheet on a named tab at the bottom

MS Excel Worksheets

- Basic worksheet mimics a relational database table
  - Rows
  - Columns
  - Column headers with names and data types
- Easy to:
  - Load from/extract CSV files
  - Load from relational database tables and queries
  - Extract to relational database tables

More Complicated Worksheets

- More complicated worksheets
  - Formatting
  - Row totals
  - Column totals
  - Calculated cells from formulae
  - Macros: mini scripts
  - Etc.
- No longer mimics a relational database table
- Needs custom code to read/write these worksheets

The End

Excel Workbooks and Worksheets

Business Cases

In-House Expert

- An in-house expert has a lot of knowledge about one specific aspect of our business.
- The expert is well-versed in MS Excel and creates many intricate workbooks.
- We need to read to extract data from these workbooks for use in our data science.
- Solution: We write processes to read the expert’s workbooks from a shared drive.

Users Who Are MS Excel Power Users

- In-house users are MS Excel power users—they have many years of using it and many years of designing workbooks.
- The output of our data science would be much more meaningful if we gave them data in workbooks in addition to reports, data visualizations, etc.
- Solution: We create MS Excel workbooks on a shared drive with the results of our data science that power users can copy and use.

The End