ETL (Extract, Transform, Load) and ELT

Concepts

ETL: Extract, Transform, Load

- Popular with traditional data warehousing
- Waterfall process
  - Each step must be completed in its entirety before next step
  - No going back—stuck with bad decisions for years
- Modern software tools and methodologies should make anyone question why anyone is still doing traditional E**T**L
- Often when someone says E**T**L they may mean E**L**T (E**T**L term was around for years before E**L**T became popular)

ELT: Extract, Load, Transform

- Popular with big data
- CI/CD (continuous integration/continuous development)
  - Each step is broken into small pieces and small pieces go through all steps
  - Continually iterate
- Modern software tools and methodologies make this the method of choice

ETL/ELT Does Not Equal Data Wrangling

- Some equate data wrangling with E**T**L/E**L**T
- E**T**L/E**L**T lacking for data science
- Culture of traditional E**T**L in data warehousing
    - Very simple data transformations
    - Relies on specifications of input datasets
    - No data exploration
    - Little or no data cleansing
    - Tends to throw out bad data even if it can be fixed
    - Little or no merging of primary datasets with secondary datasets

Waterfall

- Change is hard, expensive, long time frame
- Often gaps in requirements that take years to fix or never get fixed
- Short term saves money, long term bleeds money
- Considered obsolete in software engineering more than 20 years ago
- Still in use in some old school IT shops
- Popular with outsourcing companies
  - Maximizes billing, minimizes delivery

CI/CD

- Continually integrating
- Continually developing
- Small pieces go all the way though all steps
  - Gaps can be filled at any time
  - Agile—able to turn on a dime as our business changes
- All modern software assumes CI/CD
- Short-term costs more than waterfall—often used as an excuse to revert to waterfall

Staging Tables

Load raw data into staging tables in a relational database as quickly as possible

- Data exploration
- Data transformation
- Data cleansing

Data Exploration Using Staging Tables

- SQL
- Functional programming: procedural language and SQL
- Data exploration tools
- Tabular data charts
- Data visualizations

Concepts: ETL (Extract, Transform, Load) and ELT

The End

ETL (Extract, Transform, Load) and ELT

Business Cases

Traditional ETL Using Waterfall

- Receive a specification for dataset
- Meet with users to gather requirements for transformation
- Meet with users to finalize specifications for transformation
- Design and develop code to extract, transform, and load the dataset based on specifications
- This process takes weeks or months
- Spend all our E**T**L budget

Nasty Surprises

- Finally, after several months, we can load data
- Data is not what we expected
- Gaps: users look at the loaded data and remember some requirements they forgot
- Users are very upset and disappointed because it will not give them what they are needing
- Too much time and money to re-work; we are stuck with what we have

ELT Using CI/CD

- We receive specifications for a dataset from a vendor.
- We ask for and receive the actual data at the beginning.
- We spend a couple of days creating and loading staging tables.
- We explore the data in the staging tables and see that it’s not what we expected.
- Within days of receiving the specification, with little budget spent, we let the vendor know our concerns.

Salesman’s End Run

- Vendor sends their salesman to make an end run around IT department and tell our users that IT is lazy, does not know what they are doing, and that the data is fine.
- Salesman has a habit of buying expensive lunches and expensive gifts for our users.
- Salesman takes users to an expensive lunch and smooths over issues with their data and convinces users that data is fine.

Convince Users

- How can we convince our users that the salesman is lying?
- Using the data in the staging tables, we can quickly:
  - Find and present problem data
  - Create supporting tabular reports
  - Create supporting data visualizations
- We need to make it as easy as possible for users to understand the problems are legitimate.

Further Revisions

- Our users are convinced and support us in asking the vendor to provide data that meets our needs.
- We can repeat the process, as needed, until we get data that our users can live with.

Business Cases: ETL (Extract, Transform, Load) and ELT

The End

Data Cleansing

Concepts

Typing and Spelling Issues

- Typos
  - Berkeleu
- Misspellings
  - Berkely
- Alternate spellings
  - Sean
  - Shawn
- Garbled data
  - Drsn: Sean with hands on wrong keys

Soundex

- Old, before computers
- Phonetic
- Best uses
  - Homophones
  - Alternate spellings
  - First syllable is right

Levenshtein Distance

- Numeric measure of the difference between two strings
- Best uses
  - Typos
    - Including multiple typos
    - Including typos in the first character
  - First syllable does not have to match

Side note: Levenshtein distance can be used in machine learning and deep learning to measure string distances, including DNA strings.

Fuzzy Logic

Sophisticated combination

- Soundex
- Levenshtein distance
- Context
- Grammar
- Statistical words and phrases
  - By language, region, industry, person, etc.

Fuzzy Logic Examples

- Cisco if we are talking about IT
- Sysco if we are talking about a restaurant
- Thanks, very much apprehended
  (auto correct messed up)

Dedupe

- Data de-duplication: removing duplicates
- Sometimes legitimate
  - Two identical sales one minute apart could be customer deciding they needed another one

Missing Values

- Imputing: filling in missing values
- Possible solutions
  - Average
  - Fill down: previous value—time series, temperatures, etc.
  - Null value: leave it empty
  - Model to predict the value
- Danger: always a risk we will distort dataset

Outliers

- Keep or replace?
- Keep if we feel outliers are real exceptions that may be important
- Replace if we feel outliers are bad data

Validation Rules

- String in a date
- January 35, 2080 for a sale last week
- String in a number
- Number in a string
- 38-character last name with a 32-character column
- Temperature for Berkeley is 115 degrees
  - Outlier or validation rule?

Lookup Tables

- Sale for a product that is not in the product table
- Do we have a new product?
- Or data entry error?
- Human error

Referential Integrity

- Sale ID in the line item table that does not match a sale
- Sale ID is computer-generated
- Programming error, not human error

Consistency, Contradictions

- Sales database says a customer bought $35.49 on January 1, 2021
- Sales enterprise message queue says $37.29 on January 1, 2021

Completeness

- Sales data for store number 3 is missing on January 12, 2021
  - Store was not closed
  - Enterprise message queue has the data
- Some third-party sales channels do not include line items
  - We do not know what customers bought

Uniformity

- Some temperatures in Celsius, some in Fahrenheit
- Some sales totals include tax, some do not

Concepts: Data Cleansing

The End

Data Cleansing

Business Cases

Third-Party Sales Channels

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

Manual Entry

Some vendors have manual entry without any validation

- Typos and misspellings
  - Customer names
  - Street addresses
  - City
  - States
  - ZIP codes

Manual Entry (cont.)

Solutions

- Match against our customer database
- If we get a good match, we use our customer data
- If we do not get a good match, use fuzzy logic on names, city, states, and zip codes to see if we can find a match
- Maybe a new customer—create spreadsheet for manual review
- Unknown customer if analytical and low volume

Validation Errors

- Five-digit number in the city column
- State and ZIP are empty
- Solution
  - Probably typed the ZIP code in the city column
  - See if ZIP matches our customer table and correct based on our ZIP code table

Missing Line Items

- Some vendors do not provide us with line items
- Solutions
  - Create an unknown product
  - Try to match customer buying pattern
    - Be very careful

Contradictions

- Sales total does not match the sum of the line items
- Solutions
  - Since customer paid the sales total, assume it’s right
  - Use unknown product type to fill in missing line items
    - If low volume, may be ok
  - Maybe quantity is wrong on a line item
  - Customer buying patterns
    - Be careful

Business Cases: Data Cleansing

The End